CONTRIBUTORS TO THIS VOLUME

S. M. BLINDER
S. F. BOYS
INGA FISCHER-HJALMARS
LAURENS JANSEN
HERBERT JEHLE
PER-OLOV LÖWDIN
R. McWEENY
P. RAJAGOPAL
E. STEINER

ADVANCES IN
QUANTUM CHEMISTRY

EDITED BY

PER-OLOV LÖWDIN
DEPARTMENT OF QUANTUM CHEMISTRY
UPPSALA UNIVERSITY
UPPSALA, SWEDEN
AND
QUANTUM THEORY PROJECT
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA

VOLUME 2—1965

ACADEMIC PRESS New York • London

COPYRIGHT © 1965, BY ACADEMIC PRESS INC.
ALL RIGHTS RESERVED.
NO PART OF THIS BOOK MAY BE REPRODUCED IN ANY FORM,
BY PHOTOSTAT, MICROFILM, OR ANY OTHER MEANS, WITHOUT
WRITTEN PERMISSION FROM THE PUBLISHERS.

ACADEMIC PRESS INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS INC. (LONDON) LTD.
Berkeley Square House, London W.1

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 64-8029

PRINTED IN THE UNITED STATES OF AMERICA

LIST OF CONTRIBUTORS

Numbers in parentheses indicate the pages on which the authors' contributions begin.

S. M. BLINDER (47),[1] Department of Chemistry, Harvard University, Cambridge, Massachusetts

S. F. BOYS (1), Department of Theoretical Chemistry, University of Cambridge, Cambridge, England

INGA FISCHER-HJALMARS (25), Institute of Theoretical Physics, University of Stockholm, Stockholm, Sweden

LAURENS JANSEN (119), International Division, Battelle Memorial Institute, Geneva, Switzerland

HERBERT JEHLE (195), Physics Department, The George Washington University, Washington, D.C.

PER-OLOV LÖWDIN (213), Department of Quantum Chemistry, Uppsala University, Uppsala, Sweden, and Quantum Theory Project, University of Florida, Gainesville, Florida

R. MCWEENY (93), Quantum Theory Group, University of Keele, Keele, Staffordshire, England

P. RAJAGOPAL (1),[2] Department of Theoretical Chemistry, University of Cambridge, Cambridge, England

E. STEINER (93),[3] Quantum Theory Group, University of Keele, Keele, Staffordshire, England

[1] *Present address:* Department of Chemistry, University of Michigan, Ann Arbor, Michigan.

[2] *Present address:* Max-Plank-Institut für Physik und Astrophysik, Munich, Germany.

[3] *Present address:* Department of Chemistry, University of Exeter, England.

PREFACE

In investigating the highly different phenomena in nature, scientists have always tried to find some fundamental principles that can explain the variety from a basic unity. Today they have not only shown that all the various kinds of matter are built up from a rather limited number of atoms, but also that these atoms are constituted of a few basic elements or building blocks. It seems possible to understand the innermost structure of matter and its behavior in terms of a few elementary particles: electrons, protons, neutrons, photons, etc., and their interactions. Since these particles obey not the laws of classical physics but the rules of modern quantum theory or wave mechanics established in 1925, there has developed a new field of "quantum science" which deals with the explanation of nature on this ground.

Quantum chemistry deals particularly with the electronic structure of atoms, molecules, and crystalline matter and describes it in terms of electronic wave patterns. It uses physical and chemical insight, sophisticated mathematics, and high-speed computers to solve the wave equations and achieve its results. Its goals are great, but perhaps the new field can better boast of its conceptual framework than of its numerical accomplishments. It provides a unification of the natural sciences that was previously inconceivable, and the modern development of cellular biology shows that the life sciences are now, in turn, using the same basis. "Quantum biology" is a new field which describes the life processes and the functioning of the cell on a molecular and submolecular level.

Quantum chemistry is hence a rapidly developing field which falls between the historically established areas of mathematics, physics, chemistry, and biology. As a result there is a wide diversity of backgrounds among those interested in quantum chemistry. Since the results of the research are reported in periodicals of many different types, it has become increasingly difficult both for the expert and the nonexpert to follow the rapid development in this new borderline area.

The purpose of this serial publication is to try to present a survey of the current development of quantum chemistry as it is seen by a number of the internationally leading research workers in various countries. The authors have been invited to give their personal points of view of the subject freely and without severe space limitations. No attempts have been made to avoid overlap—on the contrary, it has seemed desirable to have certain important research areas reviewed from different points of view. The response from the authors has been so encouraging that a third volume is now being prepared.

The editor would like to thank the authors for their contributions which give an interesting picture of the current status of selected parts of quantum chemistry. Two papers dealing with problems in quantum biology are included in an attempt to start a discussion of this new important field.

It is our hope that the collection of surveys of various parts of quantum chemistry and its advances presented here will prove to be valuable and stimulating, not only to the active research workers but also to the scientists in neighboring fields of physics, chemistry, and biology, who are turning to the elementary particles and their behavior to explain the details and innermost structure of their experimental phenomena.

September, 1965 PER-OLOV LÖWDIN

Quantum Calculations

Which Are Accumulative in Accuracy, Unrestricted in Expansion Functions, and Economical in Computation

S. F. BOYS and P. RAJAGOPAL*

Department of Theoretical Chemistry
University of Cambridge
Cambridge, England

I. Introduction

The general topic which will be examined here is how far it is feasible to put into operation methods which could predict self-consistent field (SCF) solutions for molecules of moderate size to an accuracy of 1 kcal/mole in the energy. This will be referred to as chemical accuracy because it is roughly the accuracy with which it is meaningful to discuss the energy of chemical bonds and of chemical combination.

It is considered that the results of some detailed calculations which are quoted below show that we are now on the threshold of such possibilities. In these instances the methods depend on numerical integration in many dimensions. This approach has made it possible to introduce the three rather gratifying characteristics referred to in the title. The fact that they depend on the numerical integration method may not, however, be essential.

---

* Present address: Max-Planck-Institut für Physik und Astrophysik, Munich, Germany.

In the title the phrase "unrestricted in expansion functions" has been used to mean that it is feasible to use complete systems of functions such as $P(x, y, z, r) \exp(-ar)$ where $P(x, y, z, r)$ denotes any polynomial in x, y, z, r about any center in the molecule. "Economical in computation" is used to signify, in a general way, methods which are appreciably better than previous techniques. In fact, in this discussion this will be taken to imply the capacity to use functions such as $ry \exp(-ar)$ and $x^2 \exp(-ar)$ with not more than five times as much computation as previously necessary for $r \exp(-ar)$ or $y \exp(-ar)$, and a method in which the computation does not increase as fast as the squares of the number of function used.

The accumulative accuracy is a very interesting aspect which had not been completely foreseen when these investigations were begun. It implies that at any stage it is possible to restart from the best result previously determined, or from the best result of some other workers, or even with an approximation guessed from any external evidence. This aspect has not appeared to such an extent in any previous method. If previously, for example, one needed to add another expansion function to a set of ten which had already been used, then the use of the previous 2000 integrals would save some time, but in practice this is not always feasible as their accuracy might not be compatible with the calculations of the 700 or so new integrals that are required. There has been no possibility that the knowledge of the best wave function alone would enable the calculations to start from a fresh basis. In the approach discussed here this simple restart facility is possible and it is referred to as accumulative accuracy. This was in fact a characteristic of the one-dimensional integration method for atoms by the Hartree–Fock technique. At any time it was possible to start with the best results of earlier calculations and refine them further. Now it has been introduced into molecular SCF calculations, and the authors consider that it will make an appreciable contribution to the attainment of high accuracy in wave function calculations for important molecules.

In order not to overstate the potentialities these have been cited above with reference to the SCF calculations. In actual fact they apply as much to the calculations with more than one determinant in the wave function approximation, that is, the method of configurational interaction. Unfortunately it cannot be claimed that these calculations can now be performed to chemical accuracy because the new techniques do not alter the slowness of convergence as the number of detor terms is increased. This difficulty of the convergence to the true solution of Schrödinger's equation still remains, although it appears that the new techniques will contribute to the efficacy and generality of a calculation on any basis. It may be mentioned that some possibilities of using numerical integration in quantum calculations have already been demonstrated in various investigations by Frost (see Frost *et al.*, 1960; Gimarc and Frost, 1963). These authors have, however, not treated the direct evaluation of the integrals used in SCF calculations, and their methods have very little in common with the procedures discussed here.

II. The Use of Complete Systems of Functions

To state that a complete system of functions is used in a calculation does not imply that an infinite number of functions has been used but it should suggest that the functions used are the beginning of a complete system for which the use of higher terms would not introduce difficulties of a different order from those already overcome for the lower members. This requires judgment of a matter of degree, but if the next term required ten times as much effort as the last it would be misleading to describe the procedure, without qualification, as one for a complete system of functions. If it involved only four or less times the effort then it would appear very reasonable to describe it so.

In the calculation of the wave function of HCHO by Foster and Boys (1960) only functions of the Slater type for $1s$, $2s$, and $2p$ orbitals were used. The four-center integrals were evaluated by replacing $\exp(-ar)$ and $r\exp(-ar)$ by appropriate linear combinations of Gaussian functions. This procedure can be performed to any degree of accuracy. But if d functions such as $xy\exp(-ar)$ had been introduced then a very much heavier set of programs would have been required. A qualitative estimate suggests that it would be near the ten times factor. Other converging procedures for four-center integrals have been reported since then (Karplus and Shavitt, 1962), but they do not appear to make the difficulty of the d functions appreciably easier. In principle the methods are applicable to complete systems of functions but they are not in practical operation on this basis.

On the other hand, if numerical integration procedures can be put into practical operation with sufficient accuracy then the inclusion of a d function does not require much more effort than that for a p function. It might be debated that the time would be anything up to four times as great because more integration points would have to be included to obtain the integrals for d functions with the same accuracy as for the p functions. But on the other hand, if these are employed in obtaining a best fit to an orbital, then the higher functions have smaller coefficients and the same integral accuracy is not necessary. Hence it probably works out as less than two in practice. This continues just the same way for higher functions.

It is this aspect which is referred to in the title when the calculations are described as unrestricted in expansion functions. In using the programs which are in operation at present there is no feeling of an extra barrier if higher-order functions are to be used. Some of the calculations on the H_2 molecule which are quoted in Section VIII were carried out to the stage of including *s*, *p*, *d*, *f*, *g*, and *h* functions. Here *h* has been used to imply that all the relevant multinomial terms $x^5, x^4r, \ldots$ were included. The calculations in fact showed that the *g* and *h* functions were not of significance for chemical accuracy, but that the numerical operations were quite feasible.

It will be assumed here that, when the numerical integration is possible, the system of functions

$$r_I^p x_I^q y_I^s z_I^t \exp(-a_I r_I) \tag{1}$$

is the most direct one to use. The suffix I has been used to number the occupied electronic shells of the atoms with r_I the distance from the corresponding nucleus and a_I the Slater exponent for the particular shell. Such a system of functions is many times overcomplete but its early terms will give very satisfactory approximations to the wave functions. The use of overcomplete systems of functions does not create any difficulties in principle and in practice the detailed problems raised by the overcompleteness are easily solved. These difficulties are not expected to become acute since it appears that the inclusion of *d* or *f* functions will be sufficient for chemical accuracy. The same considerations would probably apply to the use of any other suitable complete system of functions and any of these could be used to any degree of accuracy if the integrals were evaluated by numerical integration.

The way in which a numerical integration procedure is valid for high- or low-order functions is easily seen by considering the three-dimensional integrals such as $\langle\varphi_1|\varphi_2\rangle$ and $\langle\varphi_1|K|\varphi_2\rangle$ where φ_1 and φ_2 could be either single terms such as $r^p x^q y^s z^t \exp(-ar)$ or linear combinations of such terms. At each point used in the numerical integration, φ_1, φ_2, and $K\varphi_2$ are evaluated. The detailed method provides a weighting factor $U(P)$ corresponding to the volume associated with this point P. Then the contributions to the total integrals are merely $U(P)\varphi_1^*(P)\varphi_2(P)$ and $U(P)\varphi_1^*(P)K\varphi_2(P)$. The value of $K\varphi_2(P) = -\frac{1}{2}\nabla^2\varphi_2(P)$ is obtained by evaluating the general formula for the derivative of each term $r^p x^q y^s z^t \exp(-ar)$ in terms of

p, q, s, t, and a and adding these together. When the integration points have been chosen by a procedure such as described in Section VI, the amount of computation varies only according to the computation required for each φ. This increases with the number of terms but not appreciably with the values of p, q, etc., for $p, q, s, t > 2$.

For six-dimensional integration of the type $(\varphi_1\varphi_1 : \varphi_2\varphi_2)$, once the combinations of points P and Q in the $\mathbf{r}_1$ and $\mathbf{r}_2$ spaces have been chosen, it is only necessary to evaluate $\varphi_1^*\varphi_1$ at P and $\varphi_2^*\varphi_2$ at Q. This integral shown in the usual notation defined in Eq. (9) is the electrostatic energy of two charge distributions, $\varphi_1^*\varphi_1$ and $\varphi_2^*\varphi_2$. In most cases all the orbitals φ that are used will be real, so it is only necessary to evaluate φ_1^2 at P and φ_2^2 at Q. Here again, the nature of the terms occurring in φ_1 and φ_2 has no significant effect on the procedure of computation. The increase in complexity in $x^3 \exp(-ar)$ compared with $\exp(-ar)$, for example, is practically negligible compared with the evaluation of r and the exponential functions. This independence of the complexity of the expansion functions is a characteristic of numerical integration but it must be noted that other possible methods show this as well. The authors at one time investigated a method depending on fitting charge distributions by linear combinations $\sum_I C_I \exp(-ar_I^2)$ and this also showed such a characteristic. But a successful general procedure was not achieved.

At the moment these procedures are only at the threshold of achieving the desired accuracy and it is quite conceivable that the detailed devices may be considerably modified in further development. Nevertheless the degree of success which has been achieved does suggest that the evaluation of all variational integrals for the higher terms is possible without any appreciable complexity or difficulty.

III. Accumulative Accuracy

At any particular stage in the procedure described below the calculation of the SCF orbitals is practically the same as the well-known procedure of using Roothaan's equations for finding the best orbitals in terms of a linear combination of n_f expansion functions $\bar{\eta}_k$ ($k = 1, 2, \ldots, n_f$). The feature which gives rise to the special characteristic of the progressive improvement is that for a $2N$ electron problem with electrons in α, β pairs the first N of the set $\bar{\eta}_k$ are the best spatial orbitals φ_k which have already been found or have been conjectured from other data. This is true even though these $\varphi_k = \bar{\eta}_k$ ($k = 1, 2, \ldots, N$) are already linear combinations from a set of n_b basis functions η_k where it is possible and very probable that

$n_b > n_f$; that is, the number of basis functions is greater than the number of degrees of freedom at the current stage.

In the earlier analytical procedures the use of n_f smaller than n_b would have involved a waste of effort because the number of integrals to be evaluated would have been $n_b^4/8$ both for $n_f < n_b$ and for $n_f = n_b$. In the first case only part of the coefficients would have been improved for roughly the same effort as is required for all the coefficients in the second case. However, with a procedure such as numerical integration where the composite integrals are evaluated directly the computing effort is less when $n_f < n_b$ than when $n_f = n_b$. To make this as clear as possible the particular case where $n_f = 2N$ will be examined. This is a reasonable practical procedure for large n_b: it corresponds to just one adjustment function for each φ orbital.

On this basis it will be shown that the computing effort for $n_f = 2N$ can be expressed in the form $16C_2n_b^2N$ while $4C_2n_b^3$ is the value for $n_f = n_b$. These are to be compared with $4C_1n_b^{4\cdot 5}$ which is derived for the analytical case in the next section. C_1 and C_2 are the appropriate unknown constants. The method which corresponds to these estimates and which has been used in practice will now be described. It should be noted also that this method has the advantage that any previous knowledge of an approximation to the φ_i can be used as an effective starting point.

The aim of the SCF calculation is to find the best determinant function

$$\Phi = \mathscr{A}(\varphi_1\alpha\varphi_1\beta\varphi_2\alpha\varphi_2\beta \cdots \varphi_N\alpha\varphi_N\beta) \tag{2}$$

which gives the lowest energy. Koopmans (1933) showed that this is achieved by taking for the φ the eigenfunctions of

$$\begin{aligned} F \equiv K + V + \sum_{j=1}^{N} \int d\mathbf{r}_2\, \varphi_j^*(\mathbf{r}_2)\varphi_j(\mathbf{r}_2)r_{12}^{-1} \\ - \sum_{j=1}^{N} \int d\mathbf{r}_2\, \varphi_j^*(\mathbf{r}_2)\varphi_j(\mathbf{r}_1)r_{12}^{-1}P(\mathbf{r}_2/\mathbf{r}_1) \end{aligned} \tag{3}$$

where $P(\mathbf{r}_2/\mathbf{r}_1)$ is defined to replace $\mathbf{r}_1$ by $\mathbf{r}_2$. The variational approximations to the solutions of this operator when

$$\varphi_i = \sum_{j=1}^{n_f} X_j^i \bar{\eta}_j \tag{4}$$

are the solutions of Roothaan's equations (1951),

$$\sum_{s}^{n_f} \{\langle\bar{\eta}_t|F|\bar{\eta}_s\rangle - E^i\langle\bar{\eta}_t|\bar{\eta}_s\rangle\}X_s^i = 0. \tag{5}$$

Since, if F is written out fully it is dependent on the X_s^i, it is only feasible to solve this by an iterative method in which the last approximation to the φ_i is used in the evaluation of F.

Let φ_i^- be the best estimates to the φ_i from data available before the current iteration. Let F^- be the operator in terms of these. Then in the accumulative procedure, the φ_i^- are used both for F^- and for the first N of the $\bar{\eta}_k$ set. The rest of the $\bar{\eta}_k$ are obtained by choosing those functions from the set η_k which are estimated as likely to be most effective in improving the φ or as those whose turn has come in taking all the η in rotation.

In the most general procedure the iteration would now be repeated with the new $\varphi_k = \sum_i X_s^i \eta_s$ as the $\bar{\eta}_k$ for $k = 1$ to N and another subset of the η_k for $\bar{\eta}_k$ for $k = N + 1$ to n_f. The experience with the H_2 molecule suggests that adequate convergence is obtained if the whole set of η_k are traversed about four times with these iterations. In practice it is unlikely that this procedure will be followed exactly because less iterations will be required with the higher η_k which only enter with very small coefficients.

We can now justify the estimates for the comparative computing efforts in the different cases. For a general value of n_f the amount of work in a single iteration must be $C_2 n_b n_f^2$ because there are n_f^2 elements in the iteration matrix and each of these is dependent on integrals in which the functions each depend on n_b elementary functions which have to be evaluated at each point of integration. Let it be assumed that it is sufficient if each subset of η_k is included in an iteration four times. In this case the total number of eigenvector iterations will be $4(n_b - N)/(n_f - N)$ because the $(n_b - N)$ adjustment functions can be divided into $(n_b - N)/(n_f - N)$ groups of $n_f - N$ functions. Hence the total effort can be written as $4C_2(n_b - N)n_b n_f^2/(n_f - N)$. In the case when n_f is taken equal to $2N$, that is, the use of a number of adjustment functions equal to that of the orbitals, this reduces to $16C_2 n_b^2 N$ for large n_b. In the case $n_f = n_b$ for large n_b, the corresponding value is $4C_2 n_b^3$.

The rough estimates of the computing effort for the cases of the analytical method, the numerical integration methods with $n_f = n_b$, and the numerical integration method with $n_f = 2N$ are thus $C_1 n_b^{4.5}$, $4C_2 n_b^3$, and $16C_2 n_b^2 N$. These are not precise estimates and in practice, for small N, it appears preferable to use $n_f = N + 6$ and then put $n_f = 2N$ for N greater than 6. However, they do indicate that there is a possibility of performing computations in which the effort increases only as n_b^2 for large values of n_b.

In the whole procedure there are two iteration processes. One is caused by Roothaan's equation which requires the repeated calculation with F in terms of the previous result and the other is caused by the necessity to repeat the use of the different subsets with $n_f < n_b$ in rotation to let the X_s^i achieve their final values. It appears to be much the best policy to mix these and go from one subset to another without letting F converge for a particular subset. The adjustment of F appears to proceed in the background while the other relaxations are being effected.

It must be noted that there is one feature which is a little regrettable. The total energy of an approximation found in one iteration is only evaluated in the next iteration. It could be calculated as a separate item but this would involve nearly as much computation as for the next iteration without providing the next adjustment. So it is most conveniently regarded as obtained in the next step. It is well-known that the total energy for some doubly occupied orbitals φ_i is

$$W = \sum_i^N \langle\varphi_i|F(\varphi_i)|\varphi_i\rangle + \sum_i^N \langle\varphi_i|K + V|\varphi_i\rangle \tag{6}$$

where the F operator is dependent on the same φ_i. If φ_i^- is used to denote the predecessor of φ_i, then the quantity nearest to the energy which can be obtained while calculating φ_i is

$$\sum_{i=1}^N \langle\varphi_i|F(\varphi_i^-)|\varphi_i\rangle + \sum_{i=1}^N \langle\varphi_i|K + V|\varphi_i\rangle$$

and this is not an exact energy approximation. However, the quantity

$$\sum_{i=1}^N \langle\varphi_i^-|F(\varphi_i^-)|\varphi_i^-\rangle = \sum_{i=1}^N \langle\bar{\eta}_i|F(\varphi_i^-)|\bar{\eta}_i\rangle \tag{7}$$

is the sum of the first N diagonal elements in the eigenvector matrix in the calculation of φ_i. Further, the quantities $\langle\bar{\eta}_i|K + V|\bar{\eta}_i\rangle$ are obtained in the construction of the $\langle\bar{\eta}_i|F(\varphi_i^-)|\bar{\eta}_i\rangle$ matrix, so that

$$W^- = \sum_{i=1}^N \langle\varphi_i^-|F(\varphi_i^-)|\varphi_i^-\rangle + \sum_{i=1}^N \langle\varphi_i^-|K + V|\varphi_i^-\rangle, \tag{8}$$

the energy of the previous approximation, is obtained as a by-product in the calculation.

It is not thought that this will be much of a disadvantage in systematic calculations because the iterations will be continued until no significant changes in the φ_i occur. This will be the most realistic test of convergence. In the last iteration to confirm that no appreciable changes in φ_i occur, it is very unlikely that any appreciable change in W will occur. It must, however, be admitted that the delay in knowing W for a particular iteration will be rather a nuisance in the early stages of an investigation.

Finally, it may be noted that at the human level there is a very helpful way in which the accumulative characteristic will assist the attainment of accurate wave functions. It frequently happens that one investigator wishes to test the results of another worker, either for direct confirmation, to test a new program, or to assess how far convergence has proceeded. In a method with such an accumulative characteristic this is very simple. It is only necessary to improve the results a stage or two further with whatever adjustment functions for the new $\bar{\eta}_k$ $(k = N + 1, N + 2, \ldots, n_f)$ which the particular investigator favors. It may even be feasible for students while gaining experience to improve valuable wave functions further in accuracy.

IV. The Time Factor in Computation

It will now be shown how for n basic expansion functions the present type of computation increases roughly as $n^{1.5}$ while for the integral expansion methods the corresponding factor is $n^{4.5}$.

It has already been explained how the labor for an SCF calculation is determined chiefly by the necessity to evaluate integrals of the type $(\varphi\varphi : \varphi\varphi)$ where $\varphi_i = \sum_{s=1}^{n} X_s^i \eta_s$. If these integrals are expanded in terms of the $(\eta\eta : \eta\eta)$ this requires the computation of more than $n^4/8$ of these. Hence the integral computation increases as n^4. In fact there are other aspects concerned with the manipulation and transformation of the $\langle\eta|F|\eta\rangle$ matrices which increase the factor to n^5 although these are not so extensive until n exceeds 20 or so. The details are a little complicated and it is simplest to designate the rate in a semisymbolic way as $n^{4.5}$. In accordance with this, if the value of n is doubled the necessary computational effort is increased by a factor of about 25. All workers in these methods have been very conscious of this.

If on the other hand the $(\varphi\varphi : \varphi\varphi)$ integrals are calculated by evaluating $\varphi^2(\mathbf{r}_1)\varphi^2(\mathbf{r}_2)r_{12}^{-1}$ at a large number of points, the only increase is in the evaluation of $\varphi_i = \sum_s^n X_s^i \eta_s$ at each point. The labor increases merely as n. The constant of proportionality is rather large for sufficient points for chemical accuracy so that for small n the labor is much greater than on the older methods. However, for large n the saving in effort will be considerable and this is the justification for the economy of computation referred to in the main title.

Here again there are manipulations and other factors as shown in the preceding section so that the total effort increases as $16C_2n^2N$. It seems, however, that this rather overestimates the effort because the fact that the higher η terms have less and less effect does in practice reduce the iterations necessary for a given accuracy, especially for large n. So just as $n^{4.5}$ is more realistic than the absolute n^5, the function $n^{1.5}$ will be taken here. This means that if the number of expansion functions is doubled the computation is increased by a factor of three or so.

Estimates referring to our present programs can only be temporary because there has not been time to explore full economies on the new procedures. But if this is borne in mind it will be interesting to know that, if $Cn^{4.5}$ is the effort on the integral expansion method, then the corresponding effort in the new procedure is approximately $C(18)^3n^{1.5}$. This means that, on the count of computation only, the old methods are best below eighteen expansion functions and the new above. It is very likely that the figure of eighteen will soon be reduced. This comparison is independent of the aspect discussed previously that the new method has been used with functions containing x^2 and x^5 while the old has not been used above x.

The most disappointing aspect is that the Ne + H_2 calculation described later required about 40 hours of computation on ESDAC 2. This, however, is only regarded as a temporary difficulty because the new machines becoming available are up to 100 times faster for some of the much used arithmetical operations. In the second place, there are possibilities of exploring fresh principles in the allocation of integration points and of straightforward improvements in the efficiency of the automatic programs.

On the other hand, for large projects, either for large molecules or large numbers of expansion functions, these new methods are probably the most effective available even at their present stage of development.

It is considered that everything indicates that the numerical integration or some other procedure in which the computation increases as $n^{1.5}$ is the most suitable for ambitious further investigations. However, the feasibility of practical projects is dominated by the shortest time in which the calculations can be performed. No appreciable survey of the merits of different schemes of point integration has yet been made. It was found possible to get only one scheme which would approach chemical accuracy. In the following sections the general features of this and the results for two systems are reported to enable the reader to make his own assessment of the future developments.

V. The Expansion of Many-Center Functions in Terms of Single-Center Functions

The most difficult integrals which have to be evaluated are of the type

$$(\varphi\varphi : \varphi\varphi) = \iint d\mathbf{r}_1\, d\mathbf{r}_2\, \varphi^2(\mathbf{r}_1)\varphi^2(\mathbf{r}_2)r_{12}^{-1} \tag{9}$$

where the φ behave like $\exp(-ar_I)$ in the neighborhood of the various nuclei. The φ^2, or $\varphi_i\varphi_j$ in the general case, are somewhat different in form from the φ_i but they show these same characteristics. The conditions are extremely unfavorable for numerical integration. The following is a device by which any function $F(\mathbf{r})$ or the $\varphi_i^2(\mathbf{r})$ type can be expanded into components each having only one cusp. For detailed reasons, it is better to make a further subdivision so that the components correspond to the electronic shells of the atoms.

The component functions, F_s, are defined by

$$F_s(\mathbf{r}) = F(\mathbf{r})V_s(\mathbf{r})/\sum_t V_t(\mathbf{r}). \tag{10}$$

It will be seen that

$$\begin{aligned} F(\mathbf{r}) &= \sum_s F(\mathbf{r})V_s(\mathbf{r})/\sum_t V_t(\mathbf{r}) \\ &= \sum_s F_s(\mathbf{r}). \end{aligned} \tag{11}$$

The $V_s(\mathbf{r})$ are to be chosen so that V_s is larger than all other V_t in the region where the orbitals of the sth electronic shell are greater than those of other shells. It is easiest to see the significance of this when there is just one shell per atom. In this case a satisfactory set of V_s functions is given by $1/r_I^4$ where r_I is the distance from the various nuclei. It is apparent that as $\mathbf{r}$ approaches the Ith nucleus, $1/r_I^4$ tends to infinity with F_I tending to F and all other F_J tending to zero. Hence none of the $F_J(J \neq I)$ has a cusp at I.

When there are two shells, as in the Ne atom, these extend to such different distances from the nucleus that it is desirable to split the component function for these regions into different parts and to integrate these with different sets of points. It can be seen that the functions V_1 and V_2 in the following set cause such a resolution. The letters A, B, and C refer to the nuclei of the Ne and the two H, respectively.

$$\begin{aligned} V_1 &= j_1/r_A^4, \\ V_2 &= (1 - j_1)/r_A^4, \\ V_3 &= 1/r_B^4, \\ V_4 &= 1/r_C^4, \end{aligned} \tag{12}$$

where

$$j_1 = [1 + \exp\{2(r_A/D - 1)\}]^{-1}. \tag{13}$$

The value 0.333 assigned to D was approximately three times the radius of the inner shell.

It will be seen that when $r_A < D$ the function V_1 is greater than all other V functions and hence F_1 tends to F as r_A is reduced. For $D < r_A \ll AB$, V_2 is the largest V function. Similarly V_3 and V_4 dominate in regions close to B and C, respectively. In this way any function $F(\mathbf{r})$ can be divided into four components each concentrated around a particular nucleus and diminishing roughly as an exponential function corresponding to the particular electronic shell. The particular choice of the weighting functions had to be made after only a limited amount of testing had been possible and the authors imagine that much better functions may be found by further investigation.

The evaluation of a general electrostatic integral can now be performed by the evaluation of sixteen integrals of the type $(F_s(\mathbf{r}) : F_t(\mathbf{r}_2))$. The latter can be regarded as an integral between single-center functions. Its evaluation is still a formidable problem but it is much simpler than the general many-center integral. These components have actually been evaluated using two grids in r, θ and ϕ around each center and the details of this are given in the next section. A special device not dependent on the particular grids of points is necessary to avoid the difficulties due to the singularity in $1/r_{12}$ and this is treated in Section VII.

It will be seen that the integrals $\langle\varphi_i|\varphi_j\rangle$, $\langle\varphi_i|r_A^{-1}|\varphi_j\rangle$, and $\langle\varphi_i|K|\varphi_j\rangle$ can be integrated by the same weighting functions at a very much simpler level since they only depend on three dimensions. The first two cases have straightforward integrands which can be evaluated at the particular grid points, but for the last integrand it is necessary to specify how the differentiation is performed. The $K\varphi_j$ factor is expressed as

$$K\varphi_j = \sum_s X_s^j(-\tfrac{1}{2}\nabla^2\eta_s).$$

Since the η_s have the form

$$r^p x^q y^t z^u \exp(-ar)$$

it is simple to derive the explicit formula for each $\nabla^2\eta_s$. By use of these the numerical value of $K\varphi_j$ is obtained for any point without there being any necessity to consider any procedure such as numerical differentiation. The automatic program performs all these operations as required.

Hence all the three-dimensional integrals can be resolved into single-center integrals

$$\int F(\mathbf{r})\,d\mathbf{r} = \sum_s \int F_s(\mathbf{r})\,d\mathbf{r}$$

and the F_s can be evaluated at any point. The actual points which were used in the integration were based on an r, θ, ϕ grid and the details of this are given in the next section.

VI. A Practical Polar Grid for Integration Around a Single Center

The aim is to integrate numerically any function $F(r, \lambda, \phi)$ for which it is known that F behaves approximately as $f(\lambda, \phi)\exp(-ar)$ for large r and that F is a smooth continuous function except at the origin. Here it is convenient for the relations below to use $\lambda = \cos\theta$ but r, θ, ϕ are just the usual polar coordinates. The element of volume of integration is $r^2 dr d\lambda d\phi$. The procedure described here is the best which was found, but the number of schemes which were tested was very small compared with the number of different arrangements which can be devised for this stage.

In order to treat the infinite range of the variable r this is expressed in terms of an auxiliary variable Q so that Q changes from 0 to 1 as r changes from 0 to ∞. An integration with respect to r can then be expressed as

$$\int_0^\infty f(r)\,dr = \int_0^1 f[r(Q)]r'(Q)\,dQ$$

$$= \frac{1}{M+1}\sum_{i=1}^{M} f\left(r\left(\frac{i}{M+1}\right)\right)r'\left(\frac{i}{M+1}\right) \tag{14}$$

where r' denotes dr/dQ.

The last formula shows the summation used for the numerical integration with respect to Q. It is the simple sum of the function of Q at M equally spaced abscissas. The contributions at $Q = 0$ and $Q = 1$ are omitted because the functions required in these problems are all zero at these values. This is a formula which is perfectly correct if the function concerned could be expanded accurately in Fourier series up to $\exp(2\pi i M Q)$. As the result of a number of one-dimensional tests, chiefly concerned with transforming the cusp of $\exp(-r)$ to a smooth zero value, the following relation between r and Q was used:

$$r = dZ^4/(Z^3 + 0.25) \qquad \text{where} \quad Z = Q^{1/2}(1 - Q)^{-1/2}. \tag{15}$$

Here d is a constant conveniently called the scale factor. The accuracy is not critically dependent on d, but a value of d about three times the mean spherical radius of the electronic function concerned gives about the optimum scale. For incidental reasons, the values of d for the particular V_s zones specified in the previous section were taken as

$$d_1 = 0.333; \qquad d_2 = 0.999; \qquad d_3 = d_4 = 2.5. \tag{16}$$

If i denotes the number of Q (or r) values in increasing order as shown in the above formula, a variable can be associated with this so that q increases by 1 for each increase of 2 in i,

$$q = \text{integral part of } (i + 1)/2.$$

Then the values of λ and ϕ taken with each value of $Q = i/(M + 1)$ are (λ_u, ϕ_v) for $u = \pm 1, \pm 2, \ldots, \pm q$, with $\phi_v = 2\pi v/j$ for $V = 1, 2, \ldots, j$, where $j = 4(q + 1)$ for three-dimensional integrals or $j = 2(q + 1)$ for six-dimensional integrals. The values of λ_u, with $\lambda_{-u} = -\lambda_u$, are defined to be the roots of the Legendre polynomial of degree $2q$. These are widely tabulated together with the associated weights B which give the correct integral values for any polynomial of degree less than $4q$. Since the ϕ values are equally spaced, the resultant weight which must be assigned to such a (λ_u, ϕ_v) point is $(2\pi B_u/j)$.

If the value $j = 4q$ had been taken for any particular i value this would have given the exact answer for an angular dependence on any spherical harmonics of degree less than $4q$. In fact $j = 2(q + 1)$ was taken as a practical compromise since $4q$ was hardly feasible with the facilities available evaluating the $Ne + H_2$ integrals. The $j = 4(q + 1)$ was used as a practical procedure of improving the interval for the three-dimensional integrals which required less time.

The result of all these operations is to give the following formula for the integral of F over all space:

$$\int_0^\infty \int_{-1}^1 \int_0^{2\pi} F(r, \lambda, \phi) r^2 \, dr \, d\lambda \, d\phi$$
$$= \sum_{i=1}^{M} \sum_{u=-q}^{q} \sum_{V=1}^{j} F\left[r\left(\frac{i}{M + 1}\right), \lambda_u, \frac{2\pi V}{j}\right] r^2 r'\left(\frac{i}{M + 1}\right) \frac{2\pi B_u}{j(M + 1)}. \tag{17}$$

The development of this grid of points followed from a number of detailed investigations which were generally based on systematic theory. The final tests for the r part of the integration were based on the simultaneous evaluation of the integrals

$$\int_0^\infty r^p \exp(-r)\, dr \qquad \text{for} \quad p = 0, 2, 4, 6.$$

Then the correctness of the orthogonality integrals for atomic s, p, and d functions was used as the general criterion of quality in the development of the (λ, ϕ) points.

The results from such grids suggest that full numerical integration can provide results to the general order of chemical accuracy, but it is apparent that very useful contributions might still be made by further investigations for better grid systems.

VII. A Treatment of the $1/r_{12}$ Singularity

If the integral for the electrostatic potential between two charge distributions $C(\mathbf{r})$ and $D(\mathbf{r})$ is evaluated by the simplest method of taking a discrete set of points $\mathbf{r}_s$ $(s = 1, 2, \ldots)$ associated with the weights U_s in three-dimensional space and using these for each distribution, a result of the following form is obtained:

$$\iint d\mathbf{r}_1\, d\mathbf{r}_2\, C(\mathbf{r}_1)D(\mathbf{r}_2)r_{12}^{-1} = \sum_s \sum_t C(\mathbf{r}_s)D(\mathbf{r}_t)U_s U_t|\mathbf{r}_s - \mathbf{r}_t|^{-1}. \tag{18}$$

It is apparent that this is unreliable because if the same points are used for both the three-dimensional spaces then infinite terms occur for $\mathbf{r}_s = \mathbf{r}_t$. If different sets of points are used then the result is very sensitive to the possibility of two points being close together.

All these difficulties are avoided by the following approach. To use a finite number of points for an integration over infinite space it is necessary to assume some knowledge about the behavior of the integrand and the most frequent form for this is an assumption that the integrand can be approximated to by a finite linear combination of some specified expansion functions, say G_k. Let us make this assumption, that $C(\mathbf{r})$ and $D(\mathbf{r})$ can be accurately approximated by expansions $C = \sum_k^N H_k G_k$. The number of functions and the discrete points are to be taken equal so that G_k can be replaced by linear combinations of these, say g_k, to satisfy

$$g_k(P_l) = \delta_{kl}. \tag{19}$$

These functions g_k are localized around the particular point to which they are related. They are unity at this point and zero at all others. Further, since the weights U_k are to be chosen to give the correct integral value for any linear combination of G or g, it follows that

$$\int g_k(r)\,dr = \sum_s U_s g_k(P_s) = U_k. \tag{20}$$

Let the further assumption be made that the distribution of g_k is approximately of the form $\exp[-a_k(r - R_k)^2]$ where the constant a_k is immediately determined by

$$(\pi/a_k)^{3/2} = \int g_k(r)\,dr = U_k. \tag{21}$$

This assumption cannot be regarded at all as precise but the assumption of other functions of $(r - R_k)^2$ leads to similar results with different constants and it will be seen below that the effect of all these is in a term which has a small effect and one which diminishes with increasing numbers of points of integration.

Let us examine the evaluation of the electrostatic integral when it is assumed that the density functions can be expanded in terms of g_k so that the coefficient of g_k will be $C(\mathbf{r}_k)$. It follows immediately that

$$\begin{aligned}\iint d\mathbf{r}_1\,d\mathbf{r}_2\,C(\mathbf{r}_1)D(\mathbf{r}_2)r_{12}^{-1} &= \sum_s\sum_t C(\mathbf{r}_s)D(\mathbf{r}_t)\iint d\mathbf{r}_1\,d\mathbf{r}_2\,g_s(\mathbf{r}_1)g_t(\mathbf{r}_2)r_{12}^{-1}\\ &= \sum_s\sum_t C(\mathbf{r}_s)D(\mathbf{r}_t)U_sU_tQ_{st}\end{aligned} \tag{22}$$

where

$$Q_{st} = \frac{\displaystyle\iint d\mathbf{r}_1\,d\mathbf{r}_2\,g_s(\mathbf{r}_1)g_t(\mathbf{r}_2)r_{12}^{-1}}{\displaystyle\int g_s(\mathbf{r}_1)\,d\mathbf{r}_1\int g_t(\mathbf{r}_2)\,d\mathbf{r}_2}. \tag{23}$$

It is immediately apparent that for $|\mathbf{r}_s - \mathbf{r}_t|$ larger than the approximate radii of the g_s, g_t functions, the Q_{st} will be equal to $|\mathbf{r}_s - \mathbf{r}_t|^{-1}$. These are the circumstances for all points except those close together. Now, for the case $s = t$, we can also obtain the value Q_{ss} very easily. From explicit formulas for integrals between Gaussians (Boys, 1950) it follows that

$$Q_{ss} = (2a/\pi)^{1/2} = 2^{1/2}/U_s^{1/3} = (U_s/2.2^{1/2})^{-1/3}. \tag{24}$$

It appeared that a reasonable approximation for Q_{st} would be to assume that

$$Q_{st} = [U_s^{1/2}U_t^{1/2}/2^{3/2} + (\mathbf{r}_s - \mathbf{r}_t)^3]^{-1/3}. \tag{25}$$

This reduces to the correct formula for the above two limiting cases and it moves over smoothly from one to the other in distances of the order of $(\pi/2a)^{1/2} = (U_s/2\sqrt{2})^{1/3}$ which is the effective radius of the function g_s. In fact, $2\sqrt{2}$ was replaced by 3 because this value is based on assumptions far more crude than this accuracy and also it can be seen that the final result is very insensitive to the value of this coefficient.

It has been convenient to expound this approximation by consideration of the Gaussian functions above, but the value of this type of Q_{st} function is not as dependent on this as might be thought. It was in fact first formulated by an *ad hoc* physical examination and was in use before the above justification was devised. The final result of this is that the following formula is used to evaluate all electrostatic integrals:

$$\iint d\mathbf{r}_1\, d\mathbf{r}_2\, C(\mathbf{r}_1)D(\mathbf{r}_2)r_{12}^{-1} = \sum_s \sum_t C(\mathbf{r}_s)D(\mathbf{r}_t)U_sU_t[r_{st}^3 + U_s^{1/2}U_t^{1/2}/3]^{-1/3}. \tag{26}$$

Here the U_s are the weights of the points in the three-dimensional grids as given by the method in Section VI, and U_t similarly for the other three-dimensional space.

VIII. Some Results of Numerical Integration Calculations

Three sets of results will be described. The first two are limited to the demonstration of the accuracy attainable in the integral values in a single stage of the whole iterative calculation of a wave function. The third comprises the results of two full accumulative calculations on H_2. These latter were performed with earlier programs in which the energy accuracy now achieved in the other results had not then been attained. They show that it is quite simple to include all the terms up to the type $z^5 \exp(-ar)$, etc., without any special complexity. In fact, it appears that the highest accuracy integrations may not be necessary to achieve good wave functions.

The first calculation is that of the energy of the simplest LCAO molecular orbital approximation to the H_2 molecule (Table I). The notations R_M^3 and R_M^6 denote three- and six-dimensional integrations, respectively, in which the numbers of values of r in the three-dimensional grids associated with these were equal to M.

Since lower accuracy had to be used in the six-dimensional integration, the value of the normalization integral $\langle\varphi|\varphi\rangle$ is given to both accuracies. The integrals for the normalized function, denoted by $\bar{\varphi}$, have been calculated by dividing by the appropriate $\langle\varphi|\varphi\rangle$ integral. The exact integrals from analytical formulas are well-known for these integrals and are given in the last column. It will be noted that the agreement for the total energy and for the separate components $\langle\bar{\varphi}|K+V|\bar{\varphi}\rangle$ and $(\bar{\varphi}\bar{\varphi}:\bar{\varphi}\bar{\varphi})$ is within 1 kcal/mole (that is 0.0016 a.u.) in each case.

TABLE I

INTEGRALS AND ENERGY OF THE SIMPLEST LCAO–MO APPROXIMATION, $\varphi=(4\pi)^{-1/2}(e^{-r_A}+e^{-r_B})$ for H_2 ($AB=1.5$)

Quantity	Integration grid	Numerical integration value	Analytical value
$\langle\varphi\|\varphi\rangle$	R_{12}^3	0.861988	0.862587
$\langle\varphi\|\varphi\rangle$	R_8^3	0.861045	0.862587
$\langle\varphi\|K\|\varphi\rangle$	R_{12}^3	0.347916	0.347619
$\langle\varphi\|V\|\varphi\rangle$	R_{12}^3	−1.349266	−1.349670
$\langle\varphi\|K+V\|\varphi\rangle$	R_{12}^3	−1.001351	−1.002051
$\langle\bar{\varphi}\|K+V\|\bar{\varphi}\rangle$	R_{12}^3	−1.161677	−1.161681
$(\varphi\varphi:\varphi\varphi)$	R_8^6	0.414988	0.416311
$(\bar{\varphi}\bar{\varphi}:\bar{\varphi}\bar{\varphi})$	R_8^6	0.559737	0.559515
E	R_8^6, R_{12}^3	−1.096950	−1.097179

This calculation was made as a direct test of the general programs for any molecular system and the fact that it could have been very much abbreviated by using the symmetry round the axis was ignored. The possibility of a hidden fitting of the point grid to give satisfactory values seems to be reasonably low, because the latest alteration of the formula for the r values had been made while working to get satisfactory orthogonality integrals for the d functions in the system Ne + H_2 described below. The agreement for the energy components as well as for the total energy also supports this, because the K integral is quite different in procedure from the electrostatic contributions. The physical system is very simple but the test is quite stringent. This is particularly so because the analytical approach of using two-center integration is ignored by the program which divides the two-center distribution into two one-center distributions as described in Section V. Hence it appears a reasonable conclusion that the numerical integration procedures are working to chemical accuracy in this case.

In Table II, the numerical integration values (R_{12}^3) are given for the orthogonality integrals for the SCF orbitals from the analytical LCAO procedure for the system of a Ne atom at a distance of 2.5 a.u. from the center of a H_2 molecule. The H_2 molecule with 1.5 a.u. for the internuclear distance forms the base of an isosceles triangle. The energy integrals for this system with the inclusion of *d* orbitals are given later. This is, of course, a rather ambitious system to use for procedure tests. But in exploratory investigations of this type, if it is possible to proceed widely at an early stage, much wasted effort on special aspects of simple systems can be avoided. These constituted particularly stringent trials and were used because the authors were particularly interested in examining the possibilities of evaluating the small differences in energy which give rise to deactivation in recombination processes.

TABLE II

NUMERICAL INTEGRATION VALUES ($R_{12}{}^3$) FOR THE ORTHOGONALITY INTEGRALS FOR SCF ORBITALS OBTAINED FROM THE ANALYTICAL PROCEDURE FOR Ne + H_2 SYSTEM

	1	2	3	4	5	6
1	1.000719					
2	0.000123	0.999801				
3	−0.000011	0.000051	1.000104			
4	0.000015	0.000018	0.000013	1.000039		
5	0.000036	−0.000016	−0.000012	−0.000054	0.999609	
6	0.000016	−0.000012	−0.000025	−0.000007	0.000101	1.000154

The orthogonality integrals which should give a unit matrix are not as good as desired but they do suggest that all the ordinary transformations of orbitals, etc., will be carried out to something approaching 0.1% accuracy.

In Table III we give the SCF eigenvalues of the orbitals for the Ne + H_2 system. The second column contains the values obtained by the previous analytical programs. The third contains the same quantities exactly except that the values of the integrals in each quantity have now been obtained by numerical integration. In column 4 the eigenvalues of the SCF matrix, starting with the same functions as first approximations but performing all operations by numerical integration, are given. In column 5 the corresponding eigenvalues are given for the case when linear combinations of a set of pseudo *d* functions on the Ne atom are included. These "*d*" functions have the same exponential factor as the *p* functions but contain the quadratic factors of the type rx, ..., x^2, xy, ... No analytical program which could calculate all the necessary quantities for column 5 is yet in operation. The three-dimensional integrals were evaluated by the R_{12}^3

grid and the six-dimensional integrals by the R_6^6 grid. The differences between the second and third columns show the errors in the numerical integration and the differences between the fourth and fifth show the effect of the "*d*" functions.

TABLE III

SCF EIGENVALUES AND ENERGY OF THE ORBITALS FOR THE Ne + H_2 SYSTEM

Number of orbital	Analytical	Equivalent numerical integration	Numerical eigenvalues with *s*, *p* functions	Numerical eigenvalues with *s*, *p*, *d* functions
1	−32.516784	−32.086107	−32.063207	−32.061947
2	−1.730893	−1.738795	−1.739087	−1.738978
3	−0.690974	−0.710145	−0.709979	−0.709952
4	−0.547102	−0.574676	−0.574628	−0.574632
5	−0.545788	−0.546145	−0.546358	−0.546163
6	−0.360726	−0.388351	−0.388291	−0.388312
Total electronic energy	−137.134472	−137.132834		

It is very desirable that the numerical integration procedures should be improved so that the second and third columns agree more closely. On the other hand, considering that this is the first direct numerical integration calculation on a system of this complexity, the agreement is quite striking. If the approximate theory of identifying the eigenvalues with the ionization potentials is assumed, then the numerical errors are less than the usual errors introduced by this approximation. Let the error in the first entry corresponding to the 1*s* orbital on the Ne be ignored because this will nearly always cancel out in ordinary chemical effects. Then the differences in the eigenvalues between the analytical and numerical bases are less than 20 kcal/mole (i.e., 0.03 a.u.). It is a very common experience that these eigenvalues are very sensitive, and almost unstable compared with the total energy value. The total energy here agrees so well that the authors can only regard it as coincidental. An optimistic, but not unreasonable assessment would be that the possible total energy errors (excluding the 1*s*) might be about 5 kcal. It may be remarked that when the procedures as developed for the NeH_2 system were applied to H_2 the error for this had dropped by a factor of 10 from the earliest calculations on H_2. On such a basis the authors take the view that these techniques, although not the present programs, are capable of giving a 1 kcal/mole accuracy for the energy of the outer electrons of molecular SCF solutions.

Some evidence on the effect of including higher orbitals is given by the results of the calculation with *d* orbitals. The circumstances are rather special because these functions only enter owing to the interaction between the Ne and the H_2. The results show that they have very little effect. In fact this is possibly more a test of the reliability of the integration procedures than a measure of how much *d* orbitals will enter in more general circumstances. However, it may be stated that in other calculations the effect of the higher degree expansion functions is generally fairly small, and decreases to about 2 kcal/mole for functions of degree two above the simple atomic orbitals. It could be inferred from this that an integration grid which is satisfactory for the LCAO calculation will be very nearly so for calculations with the final accurate SCF functions. Although this is not a precise deduction it is probably reasonably correct and is a very useful fact.

The results of two SCF calculations on the H_2 molecule are given in Table IV. These were performed with earlier programs and were much quicker than the above calculations, since the symmetry around the nuclear axis was used to reduce the integrations to five and two dimensions, respectively. In this integration procedure the simple LCAO molecular orbital function gave an error of 8 kcal/mole. The energies of all other molecular orbitals were corrected on the assumption that this was a constant error. These energy results are shown in the table and with the inclusion of *h* functions, that is x^5, r^2x^3, etc., this is within about 0.5 kcal of the best result obtained by other workers.

It will be seen that the *f*, *g*, and *h* functions have affected the results less than 1 kcal/mole, in accordance with the statements above. Even these cruder programs appear to provide a reliable basis for the investigation of orbitals in accordance with the accumulative characteristic discussed in Section III. The results can be used with the latest programs for a further refinement.

If it is accepted that the error of the deep 1*s* orbital in Ne will balance out in practical calculations, or that this may be corrected otherwise, then the conclusion that the procedure will soon be capable of an accuracy of about 1 kcal/mole appears to be justified. There is not much point in examining too closely how far the NeH_2 results fall short of this at the moment, because accumulation of experience is certain to improve the accuracy, and in addition machines which are up to hundred times faster will soon be available. Hence as far as SCF approximations are concerned it appears that these will be capable of prediction to the same degree of accuracy as it is meaningful to talk of the energy of an isolated chemical bond.

TABLE IV

THE BEST SCF ORBITAL FOR H_2 ($AB = 1.5$) IN THE FORM $(\chi_A + \chi_B)^a$

	$a = 1.0$		$a = 1.2$	
$P_i(z, r)$	X_i coefficient	Energy	X_i coefficient	Energy
1	0.689886	−1.097179	0.533679	−1.124758
r	−0.130713		0.045289	
z	0.046102	−1.128139	0.037101	−1.129968
r^2	0.062011		0.034694	
rz	−0.189291		−0.092800	
z^2	−0.073112	−1.131264	−0.041274	−1.131482
r^3	−0.003298		0.018295	
r^2z	0.017124		−0.031350	
rz^2	0.000052		−0.022403	
z^3	0.005881	−1.131501	−0.012017	−1.131926
r^4	−0.005911		−0.000855	
r^3z	0.037108		0.007100	
r^2z^2	−0.008060		0.000670	
rz^3	0.001012		0.002501	
z^4	−0.002341	−1.131907	−0.000506	−1.132288
r^5	0.000763		−0.000018	
r^4z	−0.000023		−0.000071	
r^3z^2	0.000017		0.000082	
r^2z^3	0.000019		−0.000012	
rz^4	0.000003		0.000009	
z^5	0.000047	−1.131956	−0.000011	−1.132394

[a] Here $\chi_A = e^{-ar_A} \sum_i X_i P_i(z_A, r_A)$. X_i are the coefficients of the normalized functions corresponding to the polynomicals P_i as shown. The energies are those for variational expansions up to the polynomial shown. SCF energy $= -1.131375$ (Kolos and Roothaan, 1960).

IX. Discussion

It appears certain that numerical integration procedures will be available in the future to predict the SCF wave functions of moderately complicated molecules with chemical accuracy. They will also be capable of calculating the improvements in accuracy due to the inclusion of other Slater determinants in the approximation, that is, the configurational interaction corrections. They will not, unfortunately, contribute directly to improving the slowness of the convergence of the latter.

The particular integration techniques described above suffice to show potentialities of the method but there is no reason to suppose that there are not very much better detailed methods. Probably the fastest progress will be made if various investigators can explore the most effective point systems.

In addition to the three new general characteristics which have been discussed above there is also another interesting possibility. The numerical integration method can be applied directly to give results for other types of functions. Exploratory investigations can be made with fractional powers of r used by Parr and his co-workers (Parr and Joy, 1957). They can be made for polyatomic molecules for the $\exp(-\alpha r_A - \beta r_B)$ functions shown to be effective for diatomic molecules by Harris (1960). The H_2^+ functions of Scrocco and Tomasi (1961) could also be introduced into general molecules. The value of these can now be settled by direct test.

As a conclusion to this examination of methods it is perhaps worthwhile to review what the problems are that can be solved if the techniques of full prediction from Schrödinger's equation can be advanced to a stage where they are of chemical accuracy and are in practical operation without very great labor. These can be regarded as consisting of two groups. The first is that in which experimental measurements are possible with reasonable accuracy. For these the theoretical problem is that of interpretation or diagnosis of the cause of the experimental results. The second is the class where experimental measurements are very difficult, or have not yet been made with reasonable accuracy. This classification can be summarized as follows:

(A) predictions for interpretation:
    (1) the shape and stability of molecules;
    (2) electronic spectra;
    (3) electron spin resonance and nuclear magnetic resonance;
    (4) general effects of substitution in molecules;
(B) predictions for information:
    (1) potential energy barriers in chemical reactions;
    (2) the potential fields giving deactivation in gaseous recombination processes;
    (3) the structure of radicals and unstable molecules;
    (4) the prediction of the shape and properties of molecules and radicals not yet isolated.

It is perhaps a far cry from a very wide range of predictions of these types, but it is considered that the three main characteristics which have been introduced by the numerical integration techniques discussed above will further quite considerably these aims. It might be that they can be achieved by techniques other than numerical integration but in all methods it is very desirable if we can achieve:

(1) a practicability for all types of functions;
(2) an accumulative convergence characteristic; and
(3) a computational time increasing only as a low power of the number of expansion functions.

ACKNOWLEDGEMENTS

The authors are very indebted to the Director and staff of the University Mathematical Laboratory, Cambridge, England, for the extensive facilities made available to them, and to the War Office for a grant which made the later part of these investigations possible. P. R. is grateful to the Commonwealth Scholarship Commission in the United Kingdom for a Scholarship over the period in which the investigations were commenced.

REFERENCES

Boys, S. F. (1950). *Proc. Roy. Soc.* **A200**, 542.
Foster, J. M., and Boys, S. F. (1960). *Rev. Mod. Phys.* **32**, 303.
Frost, A. A., Kellogg, R. E., and Curtis, E. C. (1960). *Rev. Mod. Phys.* **32**, 313.
Gimarc, B. M., and Frost, A. A. (1963). *J. Chem. Phys.* **39**, 1698.
Harris, F. E. (1960). *J. Chem. Phys.* **32**, 3.
Karplus, M., and Shavitt, I. (1962). *J. Chem. Phys.* **36**, 550.
Kolos, W., and Roothaan, C. C. J. (1960). *Rev. Mod. Phys.* **32**, 219.
Koopmans, T. A. (1933) *Physica* **1**, 104.
Roothaan, C. C. J. (1951). *Rev. Mod. Phys.* **23**, 69.
Parr, R. G., and Joy, H. W. (1957). *J. Chem. Phys.* **26**, 424.
Scrocco, E., and Tomasi, J. (1961). *Mol. Phys.* **4**, 193.

Zero Differential Overlap in π-Electron Theories

INGA FISCHER-HJALMARS
Institute of Theoretical Physics
University of Stockholm
Stockholm, Sweden

I. Introduction and Summary

In theoretical investigations of the electronic structure of large molecules with more or less complicated conformation complete theoretical calculations cannot be carried out at present or in the near future. All methods applied to large molecules must therefore include considerable simplifications. Especially, investigators of unsaturated systems have almost invariably assumed that it is a satisfactory approximation to study the distribution of the mobile electrons within a frame of the nuclei and the localized electrons with fixed distribution. This assumption was introduced by Hückel (1931) in his study of benzene. The theoretical background for the separation of the electronic problem in separate problems for groups of electrons with different symmetries was analyzed by Lykos and Parr (1956) and has been discussed by Lykos (1964) in an article on the π electron approximation in the previous volume of *Advances in Quantum*

Chemistry. This article also contains a broad review of the development of π-electron theories. The reader is therefore referred to Lykos' exposé and to the recent book by Parr (1963) for details about previous contributions.

The different π-electron theories show rather different degrees of sophistication. In the simplest model the nuclei and the σ electrons are supposed to give rise to a potential similar to that of a box with impenetrable walls and all the π electrons are supposed to be independent particles. Modifications leading to one-dimensional and two-dimensional potentials have also been made (see, e.g., Kuhn *et al.*, 1954, and Martin *et al.*, 1963). In the most sophisticated models the full Hamiltonian of the π electrons is considered and the π-electron distribution is determined by the Hartree–Fock method in the Linear Combination of Atomic Orbitals (LCAO) approximation. Since such a treatment implies an exceedingly time-consuming computation of a multitude of integrals, it has not yet been extended to systems with more than ten electrons. However, since machine programs for computations of many-center integrals are being developed now, one can expect new calculations of this kind to appear in the near future. Nevertheless, simpler models will remain of great value for an extensive group of molecules including the large organic molecules of vital importance in biology.

For large molecules the most frequently adopted method has been the molecular orbital (MO) method in the simple Hückel approximation. In certain investigations simplified self-consistent field methods have been used, e.g., the Hückel method including ω technique, first developed by Wheland and Mann (1949) and by Muller *et al.* (1954). Another self-consistent field method with explicit consideration of the electron repulsion terms of the Hamiltonian but formally with zero differential overlap (ZDO) has been proposed by Pariser and Parr (1953, Parr and Pariser, 1955) and by Pople (1953, 1955, 1957).

The Pariser–Parr–Pople method includes two essential simplifications in comparison to the full self-consistent LCAO-MO π-electron method. The first simplification is the ZDO assumption which implies a drastic reduction of the number of integrals, from the order of N_π^4 to the order of N_π^2, N_π being the number of π electrons in the system. The second simplification is that the values of the remaining integrals are determined in a more or less empirical way. In the present review only the first of these steps will be discussed. The choice of semiempirical parameters has been discussed many times by many different authors. Recent reviews have been given, for example, by I'Haya (1964) and Fischer-Hjalmars (1964).

Section II,A gives a brief review of the self-consistent field method.

The assumptions, typical for the ZDO approximation, are given in Eqs. (7)–(11). In Sections II,B and C various justifications for these assumptions are discussed, e.g., by application of the Mulliken approximation to the energy expressions in the MO basis and by introduction of a basis of orthogonalized atomic orbitals (OAO's). It is pointed out that such justifications are only valid in a certain approximation.

To investigate the order of this approximation in the OAO basis an expansion method is introduced in Section II,D. The expansion parameter ε is chosen to be essentially the overlap integral S_1 between adjacent atoms [Eqs. (15) and (16)]. It is clarified that the OAO's are almost as well localized as the AO's [Eq. (19)]. In Section II,E the matrix elements of the different terms in the Fock operator are expanded in the parameter ε (or S_1) [Eqs. (27)–(33)]. The matrix elements of the H^{core} operator are given in Eqs. (36) and (41). The dependence on the surrounding of the parameter W in the OAO basis (${}^{\lambda}W$) is analyzed [Eqs. (39) and (40)]. For the two-center H^{core} integral an approximate expression in terms of S_1 is derived [Eq. (46)]. The complete expression of the matrix elements of the Fock operator are given in Eqs. (52) and (53). Section II,F contains a comparison between the population analyses in the two bases, OAO and AO. It is shown that the electronic charge on the atom μ in the OAO basis, ${}^{\lambda}P_{\mu\mu}$, is approximately equal to the gross atomic population, $N_{\mu\mu}$, in the AO basis [Eqs. (57) and (58)]. An approximate relationship between the bond orders in the two bases is given in Eq. (61).

In Section III the connection of the Hückel method with the Pariser–Parr–Pople method is discussed. Two features are especially stressed, *viz.* the connection with the ω technique [Eqs. (66) and (67)] and the difference between Hückel parameters pertinent to spectral features on the one hand and to the total energy on the other hand [see Eqs. (78) and (79)]. Finally, it is pointed out that the matrix elements between nonadjacent atoms ought to be included in a consistent first-order approximation.

The conclusions in Section IV mainly stress that the ZDO computational scheme corresponds to a second-order approximation in terms of the parameter ε. Moreover, all pertinent integral values are transferable from molecule to molecule to the first order in ε, and all but ${}^{\lambda}W$ also to the second order, which is the highest that is relevant to the ZDO scheme.

II. Basis of Zero Differential Overlap

A. Assumptions in the ZDO Scheme

Since various modifications of the ZDO method have been applied with considerable success in quite a few cases it seems to be worthwhile to examine more closely the interrelations between the ZDO method and the purely theoretical method as described, for example, by Roothaan (1951, 1960).

For simplicity, the following treatment is only carried out for the case of closed-shell systems, where $N_\pi = 2n$. Then the Hamiltonian operator for a π-electron system is in atomic units:

$$H = \sum_{t}^{2n} H^{\text{core}}(t) + \sum_{t>s}^{2n} r_{st}^{-1}, \tag{1}$$

where s, t indicate the number of the electron. The one-electron operator H^{core} is:

$$H^{\text{core}} = T + U. \tag{2}$$

Here, T is the kinetic energy operator and U is the potential from all the nuclei including the hydrogens and all the σ electrons. In the Hartree–Fock approximation, where superposition of configurations is neglected, the singlet ground state eigenfunction Ψ of the operator of Eq. (1) can be written as a Slater determinant of n doubly filled one-electron molecular space orbitals $\varphi_i(t)$. According to the LCAO approximation of the MO method we have

$$\varphi_i(t) = \sum_{\mu}^{m} \chi_\mu C_{\mu i}, \qquad m \geq n \tag{3}$$

where χ_μ is an atomic π orbital (AO), centered on the atom μ, and m is the number of atoms with π orbitals. The coefficients $C_{\mu i}$ are determined by a variational procedure leading to the Hartree–Fock equations

$$\sum_{\nu=1}^{m} (F_{\mu\nu} - e_i S_{\mu\nu}) C_{\nu i} = 0, \qquad \mu = 1, 2, \ldots, m. \tag{4}$$

In Eqs. (4) the pair of subscripts, $\mu\nu$, indicates a matrix element in the basis of atomic orbitals $\{\chi\}$. $\mathbf{S}$ is the overlap matrix with the elements

$$S_{\mu\nu} = (\mu|S|\nu) = \int \chi_\mu^*(t)\chi_\nu(t)\, d\tau_t. \tag{5}$$

The operator F is the one-electron Hartree–Fock operator and the parameters e_i are the eigenvalues of F. F is composed of two parts, the core part H^{core} of Eq. (2) and the electron interaction part

$$F = H^{\text{core}} + 2J - K. \tag{6}$$

J is the one-electron Coulomb operator and K is the exchange operator with the usual definitions (see, e.g., Roothaan, 1951).

To solve Eqs. (4) the matrix elements $(\mu|H^{\text{core}}|\nu)$ and the electron interaction integrals occurring in the matrix elements $(\mu|2J - K|\nu)$ must be known. As mentioned above, the number of AO integrals, arising from the electron interaction matrix, is of the order $2n^4$. The ZDO approximation reduces this number to $2n^2$, which makes it possible to extend the self-consistent treatment to molecules with as many as twenty to thirty electrons.

The ZDO assumptions, originally suggested by Pariser and Parr (1953) and by Pople (1953), may be summarized as follows:

$$^{\lambda}(\mu|S|\nu) = \delta_{\mu\nu}, \tag{7}$$

$$^{\lambda}(\mu|H^{\text{core}}|\mu) \neq 0, \tag{8}$$

$$^{\lambda}(\mu|H^{\text{core}}|\nu) \neq 0, \qquad \text{when } \mu \text{ and } \nu \text{ are neighbors}, \tag{9}$$

$$^{\lambda}(\mu|H^{\text{core}}|\nu) = 0, \qquad \text{when } \mu \text{ and } \nu \text{ are nonneighbors}, \tag{10}$$

$$^{\lambda}[\eta(1)\kappa(2)|r_{12}^{-1}|\mu(1)\nu(2)] = {}^{\lambda}(\eta\mu|\kappa\nu) = f(R, \zeta)\delta_{\eta\mu}\delta_{\kappa\nu}. \tag{11}$$

The ZDO relation of Eq. (11) was suggested for the first time by Parr (1952) previous to the rest of the assumptions, (8)–(10). In Eqs. (7)–(11) the superscript λ indicates that the basis of the matrix elements is supposed to be different from the usual atomic orbital basis $\{\chi\}$. The notation $(\eta\mu|\kappa\nu)$ as defined in Eq. (11) is the usual notation for a two-electron repulsion integral.

B. Justification of the ZDO Assumptions by the MO Basis

As is well-known, in the usual atomic orbital basis $\{\chi\}$ some of the integrals to be neglected according to Eq. (11) are quite large. As an example, the hybrid integral $(aa|ab)$ between nearest-neighbor carbon atoms is as large as 3.5 ev. However, it was pointed out by Parr (1952) and has since been discussed in great detail (see, e.g., Parr, 1963) that the rather drastic assumption, inherent in Eq. (11), may be justified by focusing attention on the energy expression in the MO basis of Eq. (3). The numerical values of the electron interaction integrals in that basis, $(\varphi_i\varphi_j|\varphi_k\varphi_l)$, are found to be almost unchanged by the introduction of the ZDO assumptions. In the case of ethylene, for example, the value of $(\varphi_1\varphi_2|\varphi_1\varphi_2)$ is changed from 4.16 ev to 3.84 ev.

This result can be understood from a study of the formal expressions of the MO integrals in terms of the AO integrals (see, e.g., Parr, 1963; Ruedenberg, 1961). Introduction of the Mulliken approximation (Mulliken, 1949)

$$\chi_\mu \chi_\nu \approx \tfrac{1}{2} S_{\mu\nu}(\chi_\mu \chi_\mu + \chi_\nu \chi_\nu) \tag{12}$$

into the complete energy expression leads to the same result as the introduction of assumptions (7) and (11). The success of the ZDO assumptions can thus be traced back to the surprising accuracy of Eq. (12) in the case of $2p\pi$ atomic orbitals.

C. Orthogonalized Atomic Orbital Basis

In 1952, shortly after the first publication of the ZDO assumption (11), it was pointed out by Löwdin in a private communication (quoted in Fumi and Parr, 1953, and Löwdin, 1955) that orthogonalized atomic orbitals, which are defined to satisfy Eq. (7) exactly, will also approximately satisfy Eq. (11).

As is well-known, orthogonalization of a nonorthogonal complete set of functions can be performed in an infinity of different ways. A method which will preserve the local character of the atomic orbitals was suggested by Slater (1930) and applied in solid state physics by Landshoff (1936) and Wannier (1937). Löwdin (1947, 1950, 1953) has reformulated the orthonormalization procedure in a very elegant way and has given a detailed discussion of its application to both solid state and molecular problems. In a discussion of the application to molecular problems Slater (1951) suggested the name OAO (Orthogonal Atomic Orbitals) for such orbitals.

According to Löwdin (1947) the OAO's λ are defined by the equation:

$$\boldsymbol{\lambda} = \boldsymbol{\chi} \mathbf{S}^{-1/2} \tag{13}$$

where $\boldsymbol{\lambda}$ and $\boldsymbol{\chi}$ are row matrices and $\mathbf{S}^{-1/2}$ is a square matrix defined in terms of the integrals $S_{\mu\nu}$ of Eq. (5). The interrelation between the λ orbitals and the ZDO assumptions (7)–(11) has been discussed by many other authors after Löwdin (1955) and Fumi and Parr (1953). As a few examples it may be mentioned that Hall (1954) and McWeeny (1955, 1956) actually employed the λ orbitals of Eq. (13) in numerical calculations of various integrals of the kinds (8)–(11). It was found, for example, that $(aa|ab) = 3.5$ ev is reduced to $^{\lambda}(aa|ab) = -0.1$ ev, and similar results were found for other integrals. Moreover, the formalism connected with the introduction of the λ orbitals has been investigated, for example, by Parr (1960) and by Lykos (1961). The extensive investigation by Ruedenberg (1961) is also intimately connected with such problems.

When the orbital basis is subjected to the transformation (13), any one-electron operator M will be transformed as follows:

$$^{\lambda}\mathbf{M} = \mathbf{S}^{-1/2}\mathbf{M}\mathbf{S}^{-1/2} \tag{14}$$

where the superscript λ indicates that the corresponding matrix is referred to the $\{\lambda\}$ basis. Matrices with no superscript are referred to the ordinary $\{\chi\}$ basis.

Obviously, the transformation (13) to (14) will render an approximate fulfillment of the ZDO assumptions (8)–(11). It may, however, be interesting to put the following question:

(*i*) *In which approximation will the OAO basis fulfill the ZDO assumptions?*

Another problem is the following. One-center integrals in the $\{\chi\}$ basis, e.g., $(\mu|M|\mu)$ and $(\mu\mu|\mu\mu)$, are obviously completely localized. The corresponding integrals in the $\{\lambda\}$ basis, however, will at least formally depend on the whole system. We are thus led to put the following second question:

(*ii*) *In which approximation are the one-center (and the two-center) integrals independent of the surrounding, i.e., transferable from molecule to molecule?*

In some recent investigations (Fischer-Hjalmars, 1965a), the present author has discussed both these questions by means of an expansion method.

D. Expansion Method

In order to answer the above-mentioned questions the OAO representation must be considered in some detail. For the general expressions below it will be assumed that all atoms of the molecular system can be numbered in such a way that the atom number p has two (or one) nearest neighbors, numbered $p - 1$ and $p + 1$, that the next nearest neighbors are numbered $p - 2$ and $p + 2$, and so on. It is, however, easy to adapt the results obtained to cases with a more complicated molecular geometry.

My purpose is to expand the right-hand sides of Eqs. (13) and (14) in powers of a parameter $\varepsilon < 1$. In solid state physics such expansions have been made in powers of overlap integrals S_{pq} of Eq. (5) without any prescriptions about p and q (Landshoff, 1936; Wannier, 1937; Löwdin, 1947). In molecular problems an appropriate parameter choice is to assume ε to be of the order of magnitude of a typical overlap integral between nearest neighbors. First-order terms in ε should then give the "tight-binding approximation" as discussed by Ruedenberg (1961) (cf. also Löwdin, 1955; Berthier *et al.*, 1963; Leroy, 1964).

For the inclusion of higher-order terms in a systematic way, some further assumptions must be made. Since the integrals S_{pq} are roughly proportional to $\exp(-R_{pq})$, where R_{pq} is the distance between atoms p and q, and since $R_{p,p+2} \approx 2R_{p,p+1}$, it seems to be reasonable to put:

$$\begin{aligned}(p|S|p) &= 1,\\ (p|S|p+1) &= \varepsilon a_{p,p+1} + \varepsilon^2 b_{p,p+1} + O(\varepsilon^3),\\ (p|S|p+2) &= \varepsilon^2 c_{p,p+2} + O(\varepsilon^3),\\ (p|S|p+3) &= O(\varepsilon^3).\end{aligned} \tag{15}$$

Higher terms could easily be included. However, the results of importance for the present investigation are obtained with neglect of higher order than the second in ε. I will therefore refrain from the consideration of higher terms, the general expressions being rather unsurveyable when too many terms are included.

It will be shown below that it is consistent with the ZDO assumptions to neglect third-order terms. Moreover, it will be seen that, if an accuracy of at least 10% is to be kept, then the expansion parameter must be $\varepsilon < 0.4$.

This shows that the ZDO scheme can be applied to all π-electron systems, but it cannot be extended to σ electron systems in general.

The general expansion (15) is the basis for the above-mentioned analysis of the ZDO assumptions (Fischer-Hjalmars, 1965a). To make the present review of the method more easily surveyable it will be restricted to a more limited, but very important class of molecules, i.e., to hydrocarbons with approximately equidistant atoms. In that case (15) becomes:

$$\begin{aligned}(p|S|p+1) &= S_1 = \varepsilon a, \qquad a \approx 1\\ (p|S|p+2) &= S_2 = \varepsilon^2 c, \qquad c \approx 1\end{aligned} \tag{16}$$

and the general matrix element of S can be written:

$$(p|S|q) = \delta_{pq} + S_1(\delta_{p,q-1} + \delta_{p,q+1}) + S_2(\delta_{p,q-2} + \delta_{p,q+2}) + O(\varepsilon^3). \tag{17}$$

In Eq. (17) and below the subscripts p, q can take any value $1, 2, \ldots, m$. Moreover, it is assumed that every matrix element will vanish, if any subscript $p, q < 1$ or $p, q > m$.

By use of the definitions of the inverse matrix and of the square root of a matrix and only retaining ε^2 terms one finds for the general matrix element of $\mathbf{S}^{-1/2}$:

$$(p|S^{-1/2}|q) = (1 + \tfrac{3}{4}S_1^2)\delta_{pq} - \tfrac{1}{2}S_1(\delta_{p,q-1} + \delta_{p,q+1}) \\ + (\tfrac{3}{8}S_1^2 - \tfrac{1}{2}S_2)(\delta_{p,q-2} + \delta_{p,q+2}) + O(\varepsilon^3). \qquad (18)$$

From the definition (13) and from (18) it is now easy to obtain a basis function of the orthogonal set $\{\lambda\}$:

$$\lambda_p = (1 + \tfrac{3}{4}S_1^2)\chi_p - \tfrac{1}{2}S_1(\chi_{p-1} + \chi_{p+1}) + (\tfrac{3}{8}S_1^2 - \tfrac{1}{2}S_2)(\chi_{p-2} + \chi_{p+2}) + O(\varepsilon^3). \qquad (19)$$

Equation (19) shows that the coefficient of χ_p is always larger than one and that the coefficients of χ_{p-2} and χ_{p+2} are quite small, being the differences between two already small numbers. Thus, λ_p is rather well localized, although the number of nodal surfaces is larger than for the usual Slater orbital χ_p. That the local character of the AO's is retained after the transformation to the OAO's has also been pointed out recently by McWeeny (1964). This local character of the basis orbitals is of utmost importance for the second simplifying step, to be introduced *after* the ZDO approximation, i.e., the determination of the integral values by semi-empirical methods. In order that such a determination of the integral values be consistent with the whole general approach it is, of course, necessary that the formal expressions for the integrals be independent of the surrounding within the required degree of approximation.

The transformation of the matrix **M** of any one-electron operator M is obtained by insertion of Eq. (18) in Eq. (14):

$$^{\lambda}(p|M|q) = (1 + \tfrac{3}{2}S_1^2)(p|M|q) \\ - \tfrac{1}{2}S_1[(p-1|M|q) + (p+1|M|q) + (p|M|q-1) + (p|M|q+1)] \\ + \tfrac{1}{4}S_1^2[(p-1|M|q-1) + (p-1|M|q+1) + (p+1|M|q-1) \\ + (p+1|M|q+1)] + (\tfrac{3}{8}S_1^2 - \tfrac{1}{2}S_2)[(p-2|M|q) \\ + (p|M|q-2) + (p|M|q+2) + (p+2|M|q)] + O(\varepsilon^3). \qquad (20)$$

In the analysis I must also discuss the two-electron operators of Eq. (1) giving rise to integrals $(pq|st)$. It is most convenient to consider such integrals as functions of two interacting charge densities Ω_{pq} and Ω_{st}:

$$(pq|st) = \int \chi_p^*(1)\chi_q(1)\, 1/r_{12}\, \chi_s^*(2)\chi_t(2)\, d\tau_1\, d\tau_2 \\ = \int \Omega_{pq}(1)\, 1/r_{12}\, \Omega_{st}(2)\, d\tau_1\, d\tau_2 = (\Omega_{pq}|\Omega_{st}) = (\Omega_{st}|\Omega_{pq}). \qquad (21)$$

It is also convenient to introduce the square matrix $\mathbf{\Omega}$

$$\mathbf{\Omega} = \boldsymbol{\chi}^{\dagger}\boldsymbol{\chi} \tag{22}$$

with the elements

$$\Omega_{pq} = \chi_p^* \chi_q. \tag{23}$$

Since $\mathbf{S}^{-1/2}$ is hermitean, it is easily found that the transformation of $\mathbf{\Omega}$ is given by

$$^{\lambda}\mathbf{\Omega} = \boldsymbol{\lambda}^{\dagger}\boldsymbol{\lambda} = \mathbf{S}^{-1/2}\mathbf{\Omega}\mathbf{S}^{-1/2}. \tag{24}$$

Thus, Eq. (20) also gives the transformation of Ω_{pq}. By combining two such transformations, any integral $(^{\lambda}\Omega_{pq}|^{\lambda}\Omega_{st})$ can be expressed in terms of the integrals $(pq|st)$ of the AO basis. Since this expression is somewhat complicated for the general case I will confine myself to the result for special cases which will be given below.

E. Application to the Fock Operator

As mentioned above, the ZDO assumptions (7)–(11) should be analyzed by aid of an expansion in powers of ε. For this purpose Eq. (20) should be applied to the different parts of the Fock operator (6). Now, it is convenient to write H^{core} as follows:

$$H^{\text{core}} = T + \sum_{\nu} (U_{\nu} - n_{\nu} J_{\nu\nu}) \tag{25}$$

where U_{ν} denotes the potential due to the *neutral* atom ν and n_{ν} is the number of π electrons (0, 1, or 2) contributed by the atom ν. The sum over ν should be over *all* the atoms, including, e.g., hydrogen atoms.

For the correct application of Eq. (20) to the different operators it is necessary to have some means to estimate the matrix elements $(\mu|M|\nu)$ and $(\eta\mu|\kappa\nu)$ in the AO basis in terms of the expansion parameter ε.

The first term in (25) is the kinetic energy operator T. It has been shown by Ruedenberg (1961) that in a basis of Slater $2p\pi$ orbitals the two-center integral $(\mu|T|\nu)$ is to a good approximation proportional the square of the overlap integral:

$$(\mu|T|\nu) \approx S_{\mu\nu}^2 (\mu|T|\mu). \tag{26}$$

For many-center penetration integrals $(\mu|U_{\kappa}|\nu)$, in principle including both Coulomb terms and exchange terms, it is common to use the Mulliken approximation of Eq. (12). As will be discussed below [cf. Eqs. (42) and (43)] this approximation leads to an overestimate of the penetration integrals. Nevertheless it should be sufficiently accurate to give the order of magnitude correctly.

As far as the two-electron integrals $(\eta\mu|\kappa\nu)$ in the $2p\pi$ basis are concerned, the Mulliken approximation (12) is known to give surprisingly accurate results.

Introduction of Eqs. (26) and (12) into Eq. (20) leads to the following expansions in powers of ε of the various integrals in the $\{\lambda\}$ basis, where for simplicity it is assumed that all atoms with π electrons are equivalent:

$$^{\lambda}(1|T|1) = (1 + 2S_1^2)(1|T|1) + O(\varepsilon^3), \tag{27}$$

$$^{\lambda}(1|T|2) = (1|T|2) - S_1(1|T|1) + O(\varepsilon^3), \tag{28}$$

$$^{\lambda}(1|U_1|1) = (1 + \tfrac{3}{2}S_1^2)(1|U_1|1) - 2S_1(1|U_1|2) + O(\varepsilon^3), \tag{29}$$

$$^{\lambda}(1|U_2|1) = (1|U_2|1) - S_1(1|U_1|2) + \tfrac{1}{4}S_1^2(1|U_1|1) + O(\varepsilon^3), \tag{30}$$

$$^{\lambda}(1|U_{\mathrm{H}1}|1) = (1|U_{\mathrm{H}1}|1) + O(\varepsilon^3), \tag{31}$$

$$^{\lambda}(1|U_1|2) = (1|U_1|2) - \tfrac{1}{2}S_1(1|U_1|1) + O(\varepsilon^3), \tag{32}$$

$$^{\lambda}\gamma_{1n} = \gamma_{1n} + \tfrac{1}{2}S_1^2[2\gamma_{1n} - \gamma_{1,n-1} - \gamma_{1,n+1}] + O(\varepsilon^3). \tag{33}$$

In Eq. (33) the following notation is used:

$$(\mu\mu|\nu\nu) = \gamma_{\mu\nu}. \tag{34}$$

The operator $U_{\mathrm{H}1}$ in Eq. (31) refers to the hydrogen atom, bound to the carbon atom No. 1. For a comparison of Eqs. (30) and (31) the following assumption is reasonable:

$$(1|U_{\mathrm{H}1}|1) \approx (1|U_2|1). \tag{35}$$

In fact both the integrals describe the interaction between a $2p\pi$ electron and a neutral atom at comparable distances. Equation (35) is also confirmed by a numerical analysis.

The integrals (27)–(33) are the only ones that will survive in the ε^2 approximation. If one wants to go a step further and include ε^3 terms the simplicity of the ZDO scheme is immediately lost. In that approximation both penetration integrals, e.g., $(1|U_1|3)$, and two-electron integrals such as $(11|12)$, etc., must be included.

Now it is easy to write down the expansions in terms of ε of the various integrals (7)–(11). Equation (7) must obviously be fulfilled in the $\{\lambda\}$ basis to any order of ε. Equation (8) becomes

$$^{\lambda}(\mu|H^{\mathrm{core}}|\mu) = {}^{\lambda}W_{\mu} - {}^{\lambda}\gamma_{\mu\mu}(n_{\mu} - 1) - \sum_{\nu}{}' \, {}^{\lambda}\gamma_{\mu\nu}n_{\nu} + O(\varepsilon^3) \tag{36}$$

where $\sum_{\nu}'$ means the sum over all atoms $\nu \neq \mu$, and

$$ {}^{\lambda}W_{\mu} = {}^{\lambda}(\mu|T + U_{\mu} + U_{\mu-1} + U_{\mu+1} + U_{\mathrm{H}\mu}|\mu) - {}^{\lambda}\gamma_{\mu\mu}. \tag{37} $$

Let

$$ W_{\mu} = (\mu|T + U_{\mu}^{+} + U_{\mu-1} + U_{\mu+1} + U_{\mathrm{H}\mu}|\mu) \tag{38} $$

where U_{μ}^{+} denotes the potential from the unipositive atom μ. Equation (38) differs from the usual expression of W_{μ} by the inclusion of the three nearest-neighbor penetration terms. In view of Eq. (35) the sum of these three terms can be assumed to have approximately the same value for all carbon atoms. By means of Eq. (38) and Eqs. (27)–(33), Eq. (37) becomes:

$$ {}^{\lambda}W_{\mu} = W_{\mu} + n_n S_1^2[(\mu|T|\mu) + (\mu|U_{\mu}|\mu) - \tfrac{1}{2}\gamma_{\mu\mu} + \tfrac{1}{2}\gamma_{\mu,\mu+1}] + O(\varepsilon^3). \tag{39} $$

Here the generalization has been introduced that the atom μ has n_n (instead of two) nearest neighbors with π orbitals. The sum of the kinetic energy and the neutral penetration potential can be considered as the "vertical electroaffinity" of the atom μ and is thus in most cases a small quantity. If this is so, Eq. (39) can be simplified:

$$ {}^{\lambda}W_{\mu} = W_{\mu} - \tfrac{1}{2}n_n S_1^2(\gamma_{\mu\mu} - \gamma_{\mu,\mu+1}) + O(\varepsilon^3). \tag{40} $$

Equation (36) with (39) or (40) shows that the diagonal matrix element of H^{core} will have the same general form in the $\{\lambda\}$ basis as in the $\{\chi\}$ basis, although the two-electron integrals γ should be slightly modified as indicated by Eq. (33). Moreover, the "one-center" integral ${}^{\lambda}W$ is slightly changed when the number n_n of nearest neighbors with π orbitals is changed. Equations (39) and (40) show, however, that it should be an acceptable first approximation to assume ${}^{\lambda}W$ to be independent of the surrounding, but an improvement to modify ${}^{\lambda}W$ with changing n_n.

The matrix elements of H^{core} between nearest neighbors, Eq. (9), becomes

$$ {}^{\lambda}(\mu|H^{\mathrm{core}}|\mu + 1) = {}^{\lambda}(\mu|T + U_{\mu} + U_{\mu+1}|\mu + 1) + O(\varepsilon^3) \tag{41} $$

since all hybrid integrals will vanish. Equation (41) emphasizes the importance of the penetration terms, being by no means negligible but of the same order of magnitude as the only other nonvanishing term. In view of Eqs. (28) and (32) it is clear that the expression (41) has a completely *local* character also in the $\{\lambda\}$ basis.

The matrix element ${}^{\lambda}(\mu|H^{\mathrm{core}}|\nu)$, where the atoms μ and ν are non-neighbors, will only contain terms of the order ε^3. Thus Eq. (10) is fulfilled to the order ε^2.

For the homonuclear case it is possible to reformulate Eqs. (28), (32), and (41) in an interesting way. The Coulomb part of the penetration integral, U^{Coul}, is found to satisfy the following approximate relations:

$$(1|U_2^{\mathrm{Coul}}|1) \approx (1|U_1^{\mathrm{Coul}}|1)S_1^2(1 + 2S_1)/3, \tag{42}$$

$$(1|U_1^{\mathrm{Coul}}|2) \approx (1|U_1^{\mathrm{Coul}}|1)S_1^2(1 - S_1). \tag{43}$$

By the aid of Eqs. (26), (42), and (43), Eqs. (28) and (32) can be written:

$$^{\lambda}(1|T|2) \approx -(1|T|2)(1 - S_1)/S_1, \tag{44}$$

$$^{\lambda}(1|U_1^{\mathrm{Coul}}|2) \approx -(1|U_1^{\mathrm{Coul}}|2)(1 - S_1)/(2S_1). \tag{45}$$

Equations (44) and (45) show that these matrix elements of the kinetic and of the penetration energy will change sign in the $\{\lambda\}$ basis as compared to the $\{\chi\}$ basis.

Moreover, since the exchange penetration is rather small in the two-center case (but not in the one-center), Eqs. (42)–(45) can be used to reformulate Eq. (41):

$$\begin{aligned} ^{\lambda}(1|H^{\mathrm{core}}|2) &\approx -(1|T + U_1|2)(1 - S_1)/S_1 \\ &\approx -(1|T + U_1^{\mathrm{Coul}}|1)S_1(1 - S_1) \\ &= \mathrm{const}\; S_1(1 - S_1). \end{aligned} \tag{46}$$

It is interesting that to the first order in S Eq. (46) gives the Mulliken relation between the semiempirical parameter β and S, and to the second order the relation suggested by Ruedenberg (see Parr, 1963, p. 100) is obtained. However, the relation is not very accurate. It should also be stressed that $(1|T + U^{\mathrm{Coul}}|1)$ is *not* equal to the ionization potential or to W, since W contains the potential U^+ from the positive ion and, of course, includes the exchange terms (cf. also Berthier *et al.*, 1963).

As is well-known, the matrix elements of the complete Fock operator (6) are most conveniently expressed by means of the charge and bond order matrix **P**, defined by Coulson (1939; Coulson and Longuet-Higgins, 1947) for the case of an orthonormalized basis:

$$^{\lambda}P_{\mu\nu} = 2\sum_{i=1}^{n} {}^{\lambda}C_{\mu i}\, {}^{\lambda}C_{i\nu}^{*} \tag{47}$$

or

$$^{\lambda}\mathbf{P} = 2\, {}^{\lambda}\mathbf{C}\, {}^{\lambda}\mathbf{C}^{\dagger}. \tag{48}$$

The relation between the coefficients $^{\lambda}C_{\mu i}$ in Eqs. (47) and (48) and $C_{\mu i}$ in Eq. (3) is obviously

$$^{\lambda}\mathbf{C} = \mathbf{S}^{1/2}\mathbf{C}. \qquad 49)$$

Considering Eq. (11) the electron interaction part of Eq. (6) can be written

$$^{\lambda}(\mu|2J-K|\nu) = \sum_{\eta=1}^{m} \sum_{\kappa=1}^{m} {}^{\lambda}P_{\eta\kappa}({}^{\lambda}\gamma_{\kappa\mu}\delta_{\eta\kappa}\delta_{\mu\nu} - \tfrac{1}{2}\,{}^{\lambda}\gamma_{\mu\nu}\delta_{\eta\mu}\delta_{\kappa\nu}). \tag{50}$$

By introduction of the net atomic charge, originating from the π electron distribution

$${}^{\lambda}Q_{\mu} = n_{\mu} - {}^{\lambda}P_{\mu\mu} \tag{51}$$

the matrix elements of the Fock operator can be written

$$^{\lambda}(\mu|F|\mu) = {}^{\lambda}W_{\mu} + {}^{\lambda}\gamma_{\mu\mu}(1 - \tfrac{1}{2}n_{\mu}) - \tfrac{1}{2}\,{}^{\lambda}Q_{\mu}\,{}^{\lambda}\gamma_{\mu\mu} - \sum_{\nu}{}' \,{}^{\lambda}Q_{\nu}\,{}^{\lambda}\gamma_{\mu\nu}, \tag{52}$$

$$^{\lambda}(\mu|F|\nu) = {}^{\lambda}(\mu|H^{\text{core}}|\nu) - \tfrac{1}{2}\,{}^{\lambda}\gamma_{\mu\nu}\,{}^{\lambda}P_{\mu\nu}. \tag{53}$$

In (53) the first term on the right-hand side will disappear when $\nu \neq \mu + 1$.

F. Population Analysis

From the self-consistent treatment of the Hartree–Fock equations (4) the coefficients ${}^{\lambda}\mathbf{C}$ and the charge and bond order matrix ${}^{\lambda}\mathbf{P}$ are determined. It is usual to interpret ${}^{\lambda}P_{\mu\mu}$ as the π-electron charge on the atom μ and ${}^{\lambda}P_{\mu\nu}$ as the bond order between the atoms μ and ν. As has been pointed out by Peacock (1959) ${}^{\lambda}\mathbf{P}$ differs from the matrix $\mathbf{P}$ which is connected with the usual atomic orbital basis $\{\chi\}$. In fact

$$\mathbf{P} = \mathbf{S}^{-1/2}\,{}^{\lambda}\mathbf{P}\mathbf{S}^{-1/2}, \tag{54}$$

as is easily seen from Eqs. (48) and (49). Peacock has also compared numerical values of the matrix elements of $\mathbf{P}$ and of ${}^{\lambda}\mathbf{P}$ for a series of molecules and obtained appreciable differences.

It must be noticed, however, that the diagonal elements ${}^{\lambda}P_{\mu\mu}$ are not to be compared with the elements $P_{\mu\mu}$, which are equal to the net atomic populations (Mulliken, 1955), but with the gross atomic populations $N_{\mu\mu}$:

$$N_{\mu\mu} = \sum_{\nu} P_{\mu\nu}S_{\nu\mu} \tag{55}$$

which are the diagonal elements of a matrix $\mathbf{N}$. Equations (54) and (55) give:

$$\mathbf{N} = \mathbf{S}^{-1/2}\,{}^{\lambda}\mathbf{P}\mathbf{S}^{1/2} \tag{56}$$

and

$$N_{\mu\mu} = {}^{\lambda}P_{\mu\mu}, \qquad \text{if } [\mathbf{S}, {}^{\lambda}\mathbf{P}] = 0. \tag{57}$$

If $\mathbf{S}$ and ${}^{\lambda}\mathbf{P}$ do not commute, it can be shown by use of the expansion method that

$$N_{\mu\mu} = {}^{\lambda}P_{\mu\mu} + O(\varepsilon^2), \qquad \text{if } [\mathbf{S}, {}^{\lambda}\mathbf{P}] \neq 0. \tag{58}$$

For the net atomic charge Q, one finds from Eqs. (51), (57), and (58),

$$^{\lambda}Q_{\mu} = Q_{\mu} + O(\varepsilon^2) \tag{59}$$

where $O(\varepsilon^2)$ vanishes if $\mathbf{S}$ and $^{\lambda}\mathbf{P}$ commute. The conclusion is that the common interpretation of $^{\lambda}P_{\mu\mu}$ and $^{\lambda}Q_{\mu}$ is correct at least to the first order, and in most cases also to the second order. This is obviously the best that can be expected for a ZDO method.

It is not so easy to compare bond orders $^{\lambda}P_{\mu\nu}$ and overlap populations $n_{\mu\nu}$ (cf., however, Ruedenberg 1958):

$$n_{\mu\nu} = 2P_{\mu\nu}S_{\nu\mu}. \tag{60}$$

When only first-order terms are included, one obtains

$$P_{\mu\nu} = {}^{\lambda}P_{\mu\nu} - \tfrac{1}{2}S_{\mu\nu}[{}^{\lambda}P_{\mu\mu} + {}^{\lambda}P_{\nu\nu}] + O(\varepsilon^2) \approx {}^{\lambda}P_{\mu\nu} - S_{\mu\nu}. \tag{61}$$

The expression including second-order terms also can easily be obtained by use of Eqs. (54) and (20). The general trend is, however, obvious already from the simple expression (61). It shows that in most cases the sequence of the different bond orders will be the same in the $\{\lambda\}$ basis as in the $\{\chi\}$ basis. However, when two bond orders are almost equal the sequence may be inverted.

III. The Hückel Method

A. Connection Problem

The problem of the rational connection of the Hückel method with the Pariser–Parr–Pople method has been discussed in great detail previously, e.g., by Pople (1953), McWeeny (1955, 1964), and Del Re and Parr (1963). In the following section two features of this problem will be especially stressed, *viz.*: (*i*) the connection of the ω technique with the Pariser–Parr–Pople method; (*ii*) the important difference in the connection problem for applications to spectra and to ground state properties.

The leading thought of the following discussion is the reasonable assumption that the Hückel method has to be interpreted as an expansion in ε, including first-order terms only.

In the present article, the letters α and β are used only to denote the semiempirical parameters in the Hückel method.

As is well-known, in the Hückel approximation the electronic structure is determined by the solutions to the equations

$$(\alpha_{\mu} - e_i')C_{\mu i} + \sum_{\nu}{}' \beta_{\mu\nu}C_{\nu i} = 0, \qquad \mu = 1, 2, \ldots, m \tag{62}$$

where the sum over ν is restricted to the nearest neighbors of the atom μ. A comparison of Eqs. (62) and (4) indicates that a connection between the Hückel method and the self-consistent field method could be obtained by putting

$$\begin{aligned} \alpha_\mu &= F_{\mu\mu}, \\ \beta_{\mu\nu} &= F_{\mu\nu} \quad \text{when } \mu \text{ and } \nu \text{ are neighbors,} \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{63}$$

If it is possible to choose semiempirical α and β values that are reasonably good approximations to the self-consistent matrix elements of the Fock operator, the solution of the simple Hückel equations (62) should be sufficient for the prediction of such observables, which depend mainly on the eigenvalues and the eigenvectors of the Fock operator. Such observables are, for example, the ionization potential and the spectral frequency and intensity (disregarding the degeneracy problem). These observables are obviously *differences* between certain quantities pertinent to different electronic states of a fixed framework.

On the other hand, when the *total* quantities (e.g., total energy) are of importance, or the difference between quantities, pertinent to electronic states of different frameworks (e.g., resonance energy), the Hückel parameters must be chosen in another way than indicated by Eq. (63).

B. The ω Technique

For the case of alternant hydrocarbons Pople (1953) has shown the following relation between α of the Hückel approximation and the various integrals of the ZDO approximation:

$$\alpha_C^0 = W_C + \tfrac{1}{2}\gamma_{CC}. \tag{64}$$

The superscript 0 on α indicates that this expression refers to the case with zero charge on the atom.

From Eq. (52) it is seen that, since W_μ and $\gamma_{\mu\mu}$ to the first order in ε are the same in the $\{\lambda\}$ basis and the $\{\chi\}$ basis, a more general expression should be

$$\alpha_\mu^0 = W_\mu + \gamma_{\mu\mu}(1 - \tfrac{1}{2}n_\mu). \tag{65}$$

Moreover, Eq. (52) also shows that a net charge Q_μ [cf. Eq. (59)] on the atom μ should be taken into account by the following adjustment of the α value:

$$\alpha_\mu = \alpha_\mu^0 - \tfrac{1}{2}Q_\mu\gamma_{\mu\mu}. \tag{66}$$

As is well-known, in the ω technique the following relationship is used (cf., e.g., Streitwieser, 1961, where further references are given):

$$\alpha_\mu = \alpha^0_\mu + \omega\beta^0 Q_\mu. \tag{67}$$

A comparison of Eqs. (66) and (67) shows that $\omega = -\gamma_{\mu\mu}/(2\,\beta^0)$, giving $\omega \approx 2.4$. This value is somewhat higher than those used in recent calculations, in line with the obvious fact that a theoretical ω value must be too high due to neglect of polarization of the σ core.

Equation (52) also indicates a possible modification of α_μ when there are considerable charges Q_ν on the neighboring atoms. However, such an elaboration of the Hückel method does not seem to be consistent with the substantial approximation to consider electron interactions only indirectly.

C. The Total Energy

As has been discussed in great detail by Mulliken (1949), the numerical value of the parameter β turns out to be rather different when determined by means of different empirical quantities. Thus Mulliken distinguishes, for example, between β^{spectr} and β^{ad}, determined from spectral energies and from dissociation energies, respectively. This problem has recently been reconsidered by several authors, e.g., McWeeny (1964), who introduced the superscript "ad" for "additive partition." McWeeny has shown that β should be identified with different ZDO expressions in the two cases:

$$\beta^{\text{spectr}}_{\mu\nu} = {}^\lambda(\mu|F|\nu), \tag{68}$$

$$\beta^{\text{ad}}_{\mu\nu} = {}^\lambda(\mu|F|\nu) + \tfrac{1}{4}\gamma_{\mu\nu}\,{}^\lambda P_{\mu\nu}. \tag{69}$$

In (69) it is understood that μ and ν are neighboring atoms. Ruedenberg (1961) and Del Re and Parr (1963) have also given detailed discussions of this problem. The following derivation is slightly different from those published previously but most closely related to the much more elaborate and complete presentations of Ruedenberg and of Del Re and Parr.

In the self-consistent field case the total π electron energy E_π can be written

$$E_\pi = \tfrac{1}{2}\,\mathrm{Tr}\{\mathbf{P}(\mathbf{H}^{\text{core}} + \mathbf{F})\}. \tag{70}$$

The total energy of the system, E_{tot}, is the sum of E_π and the interaction between the positively charged core atoms, E_{repuls}, which can be written (cf. e.g., Del Re and Parr, 1963):

$$E_{\text{repuls}} = \tfrac{1}{2}\sum_\mu \sum_\nu{}' n_\mu n_\nu \gamma_{\mu\nu}, \tag{71}$$

$$E_{\text{tot}} = E_\pi + E_{\text{repuls}}. \tag{72}$$

Using Eqs. (25), (50), (70), and the following notation,

$$\begin{aligned} f_{\mu\mu} &= W_\mu + \gamma_{\mu\mu}(1 - \tfrac{3}{4}n_\mu) - \tfrac{1}{4}\gamma_{\mu\mu}Q_\mu, \\ f_{\mu\nu} &= {}^\lambda(\mu|H^{\text{core}}|\nu) - \tfrac{1}{4}\gamma_{\mu\nu}P'_{\mu\nu}, \end{aligned} \tag{74}$$

with

$$P'_{\mu\nu} = {}^\lambda P_{\mu\nu} - 2Q_\mu Q_\nu/{}^\lambda P_{\mu\nu}, \tag{75}$$

after some straightforward algebra the total energy (72) can be written (Fischer-Hjalmars, 1965b):

$$E_{\text{tot}} = \text{Tr}\{{}^\lambda\mathbf{Pf}\}. \tag{76}$$

In the Hückel theory, the energy is

$$E_{\text{Hückel}} = 2\sum_i e'_i = \sum_\mu \sum_\nu {}^{\text{H}}P_{\mu\nu}\beta_{\mu\nu} = \text{Tr}\{{}^{\text{H}}\mathbf{P}\boldsymbol{\beta}\} \tag{77}$$

where ${}^{\text{H}}\mathbf{P}$ is the charge and bond order matrix from Eqs. (62) and $\beta_{\mu\mu} = \alpha_\mu$. Obviously, the additivity of the Hückel energy is maintained if

$$\begin{aligned} \alpha_\mu^{\text{ad}} &= f_{\mu\mu} = \alpha_\mu^{\text{spectr}} - \tfrac{1}{4}\gamma_{\mu\mu}\,{}^\lambda P_{\mu\mu} + \sum_\nu{}' \,Q_\nu\gamma_{\mu\nu}, \\ \beta_{\mu\nu}^{\text{ad}} &= f_{\mu\nu} = \beta_{\mu\nu}^{\text{spectr}} + \tfrac{1}{4}\gamma_{\mu\nu}[{}^\lambda P_{\mu\nu} + 2Q_\mu Q_\nu/{}^\lambda P_{\mu\nu}], \end{aligned} \tag{78}$$

where [cf. Eqs. (52) and (53)]

$$\begin{aligned} \alpha_\mu^{\text{spectr}} &= {}^\lambda(\mu|F|\mu), \\ \beta_{\mu\nu}^{\text{spectr}} &= {}^\lambda(\mu|F|\nu). \end{aligned} \tag{79}$$

The difference between the present values of α^{ad} and β^{ad} and those of McWeeny is that E_{repuls} has been included here, as has been done by Pople (1953), and by Del Re and Parr (1963). Although Pople never derived any expressions for α^{ad} and β^{ad}, expressions corresponding to Eqs. (78) can be obtained from his equation by regrouping the terms. It seems to be in line with Hückel's original philosophy to include E_{repuls}. Thus it becomes clear that only the small net atomic charges Q_μ contribute to the total energy.

The "bond order" $P'_{\mu\nu}$ of Eq. (75) can be interpreted in an interesting way. The original bond order ${}^\lambda P_{\mu\nu}$ becomes modified by the net charges on the two atoms, either increased, when the two charges have different signs, or diminished, when the signs are equal.

D. Matrix Elements between Nonneighbors

Both the Hückel parameters $\beta_{\mu\nu}^{\text{spectr}} = {}^\lambda(\mu|F|\nu)$ and $\beta_{\mu\nu}^{\text{ad}}$ of Eqs. (53), (79), and (74) are usually assumed to vanish when μ and ν are nonneighbors. However, as was mentioned above, the Hückel approximation should reasonably correspond to a ZDO approximation including all the first-order terms. In this approximation the matrix elements of H^{core} between nonneighbors will certainly vanish, but $-\frac{1}{2}P_{\mu\nu}\gamma_{\mu\nu}$ in Eq. (53) will still remain. It is usually assumed that these terms are small. Some arguments to support this assumption have been given by Murrell and Salem (1961). Actual applications also show that the result will be approximately the same whether these terms are included or not.

Recently, a new semiempirical approach has been proposed by Orloff and Fitts (1963), which only differs from the Hückel method by the inclusion of all the off-diagonal matrix elements. Since these authors want to retain the simplicity of the Hückel method, they suggest that the matrix elements between nonneighbors be estimated from previous self-consistent calculations in the ZDO scheme for related molecules. Another possibility seems to be to start with a Hückel calculation, giving approximate values of all the $P_{\mu\nu}$ which can be used for the construction of the nonneighbor matrix elements. Afterwards, a calculation according to the scheme by Orloff and Fitts can easily be made. Thus this scheme offers a consistent method to include all the terms pertinent to the first-order approximation.

IV. Concluding Remarks

As already mentioned, the new development of machine programs for computation of integrals, which will make possible more extensive *ab initio* calculations than have been feasible hitherto, is not likely to make simplified models superfluous for many years to come. It is therefore of great interest to put the footing of such models on as firm a ground as possible and to suggest improvements in a consistent way. The above analysis of the ZDO approximations aims at a contribution along such lines.

The analysis has shown that the ZDO approximations are equivalent to an expansion in terms of a parameter ε of the OAO representation of the Fock equations, *including at least first-order and at most second-order terms in* ε. First-order terms must be included, since otherwise all the off-diagonal matrix elements of H^{core} will disappear [cf. Eqs. (41), (28), (32), or (46)]. Third-order terms cannot be included, since, for example, all the hybrid integrals are of this order. Hence, the ε^3 terms invalidate Eq. (11). Since all matrix elements of H^{core} between nonneighbors are of third order or smaller, it is inconsistent to include such terms, but still keep to Eq. (11), as has sometimes been suggested.

Since third-order terms will invalidate the ZDO scheme, the method should only be applied to such problems where moderate accuracy can lead to interesting information. When very high accuracy is necessary, one must obviously turn to much more elaborate methods.

The Pariser–Parr–Pople method includes both the ZDO approximations and the determination of certain integral values in an empirical way. The method has been criticized for inconsistency in this respect. It has been pointed out that the ZDO assumptions can only be justified by an orthonormalized basis, but that this very basis will destroy the local character of the various integrals and therefore prohibit the transference of numerical values from one molecule to another. However, the above analysis has shown that *to the first order in* ε, *all the pertinent integrals are independent of the surrounding* [cf. Eqs. (27)–(33), (40), and (41)]. A closer examination shows that when second-order terms are included all integrals except ${}^{\lambda}W$ remain transferable (Fischer-Hjalmars, 1965a).

It has been discussed whether the ZDO approximations (7)–(11) or a consistent use of the Mulliken approximation (12) should be preferable. In many respects these two approaches are connected. However, when semiempirical integral values are introduced, a bookkeeping must be accomplished between, for example, γ values arising on the one hand from the original energy expression and on the other hand from the application of the Mulliken approximation. To the former integrals empirical values, γ_{emp}, may be assigned while the latter should be treated as theoretical quantities, γ_{theor} [cf., e.g., I'Haya (1964)]. This double bookkeeping will certainly complicate the Mulliken method. It is avoided in the ZDO method since the hybrid and exchange integral will vanish in the ε^2 approximation.

The relation between the Hückel method and the ZDO method indicates that when an appropriate version of the Hückel method is applied *either* to such ground state properties as resonance energy, etc., *or* to spectral features rather good results may be expected. Still better results may be obtained by the extension of the Hückel method suggested by Orloff and Fitts (1963).

Whether an actual problem should be treated by the Hückel method, by the Orloff and Fitts method, or by the ZDO method depends, of course, on the complexity of the problem and the desired accuracy. All three methods seem to have their important fields of application. Moreover, experience gained from one of the methods can be of great value for the application of the other methods.

Attempts have been made to extend the computational scheme of the ZDO method to σ orbitals. In principle, such an extension may be successful when the overlap integrals remain sufficiently small. As an example, when $S = 0.1$ the integral $(\sigma_1\sigma_1|\sigma_1\sigma_2)$ is 1.5 ev in the $\{\chi\}$ basis but reduced to -0.1 ev in the ε^2 approximation of the $\{\lambda\}$ basis. However, in most cases the σ overlap integrals are quite large, often greater than 0.4. In such cases the results from ZDO computations must be interpreted with much reservation.

REFERENCES

BERTHIER, G., BAUDET, J., and SUARD, M. (1963). *Tetrahedron* **19** Suppl. 2, 1.

COULSON, C. A. (1939). *Proc. Roy. Soc.* (*London*) **A169**, 413.

COULSON, C. A., and LONGUET-HIGGINS, H. C. (1947). *Proc. Roy. Soc.* (*London*) **A191**, 39.

DEL RE, G., and PARR, R. G. (1963). *Rev. Mod. Phys.* **35**, 604.

FISCHER-HJALMARS, I. (1964). *In* "Molecular Orbitals in Chemistry, Physics, and Biology" (P.-O. Löwdin and B. Pullman, eds.), p.361. Academic Press, New York.

FISCHER-HJALMARS, I. (1965a). *J. Chem. Phys.* **42**, 1962.

FISCHER-HJALMARS, I. (1965b). To be published.

FUMI, F. G., and PARR, R. G. (1953). *J. Chem. Phys.* **21**, 1864.

HALL, G. G. (1954). *Trans. Faraday Soc.* **50**, 773.

HÜCKEL, E. (1931). *Z. Physik* **70**, 204.

I'HAYA, Y. (1964). *Advan. Quant. Chem.* **1**, 203.

KUHN, H., HUBER, W., and ENGELSTÄTTER, H. (1954). *Z. Phys. Chem.* **1**, 142.

LANDSHOFF, R. (1936). *Z. Physik* **102**, 201.

LEROY, G. (1964). *Bull. Soc. Chim. Belges* **73**, 166.

LYKOS, P. G. (1961). *J. Chem. Phys.* **35**, 1249.

LYKOS, P. G. (1964). *Advan. Quant. Chem.* **1**, 171.

LYKOS, P. G., and PARR, R. G. (1956). *J. Chem. Phys.* **24**, 1166; **25**, 1301.

LÖWDIN, P.-O. (1947). *Arkiv Matem. Astron. Fysik* **35A**, No. 9.

LÖWDIN, P.-O. (1950). *J. Chem. Phys.* **18**, 365 (1950).

LÖWDIN, P.-O. (1953). *J. Chem. Phys.* **21**, 496.

LÖWDIN, P.-O. (1955). *Svensk Kem. Tidskr.* **67**, 380.

MARTIN, H., FÖRSTERLING, H. D., and KUHN, H. (1963). *Tetrahedron* **19** Suppl. 2, 243.

MCWEENY, R. (1955). *Proc. Roy. Soc.* (*London*) **A227**, 288.

MCWEENY, R. (1956). *Proc. Roy. Soc.* (*London*) **A237**, 355.

MCWEENY, R. (1964). *In* "Molecular Orbitals in Chemistry, Physics, and Biology" (P.-O. Löwdin and B. Pullman, eds.), p.305. Academic Press, New York.

MULLER, N., PICKETT, L. W., and MULLIKEN, R. S. (1954). *J. Am. Chem. Soc.* **76**, 4770.

MULLIKEN, R. S. (1949). *J. chim. phys.* **46**, 497 and 675.

MULLIKEN, R. S. (1955). *J. Chem. Phys.* **23**, 1833.

MURRELL, J. N., and SALEM, L. (1961). *J. Chem. Phys.* **34**, 1914.

ORLOFF, M. K., and FITTS, D. D. (1963). *J. Am. Chem. Soc.* **85**, 3721.

PARISER, R., and PARR, R. G. (1953). *J. Chem. Phys.* **21**, 466 and 767.

PARR, R. G. (1952). *J. Chem. Phys.* **20**, 1499.

PARR, R. G. (1960). *J. Chem. Phys.* **33**, 1184.

PARR, R. G. (1963). "Quantum Theory of Molecular Electronic Structure." Benjamin, New York.

PARR, R. G., and PARISER, R. (1955). *J. Chem. Phys.* **23**, 711.
PEACOCK, T. E. (1959). *J. Chem. Soc.*, p.3241.
POPLE, J. A. (1953). *Trans. Faraday Soc.* **49**, 1375.
POPLE, J. A. (1955). *Proc. Phys. Soc. (London)* **A68**, 81.
POPLE, J. A. (1957). *J. Phys. Chem.* **61**, 6.
ROOTHAAN, C. C. J. (1951). *Rev. Mod. Phys.* **23**, 69.
ROOTHAAN, C. C. J. (1960). *Rev. Mod. Phys.* **32**, 179.
RUEDENBERG, K. (1958). *J. Chem. Phys.* **29**, 1232.
RUEDENBERG, K. (1961). *J. Chem. Phys.* **34**, 1861, 1878, 1884, 1892, 1897, 1907.
SLATER, J. C. (1930). *Phys. Rev.* **35**, 509.
SLATER, J. C. (1951). *J. Chem. Phys.* **19**, 220.
STREITWIESER, A., JR. (1961). "Molecular Orbital Theory for Organic Chemists." Wiley, New York.
WANNIER, G. H. (1937). *Phys. Rev.* **52**, 191.
WHELAND, G. W., and MANN, D. E. (1949). *J. Chem. Phys.* **17**, 264.

Theory of Atomic Hyperfine Structure

S. M. BLINDER*

Department of Chemistry
Harvard University
Cambridge, Massachusetts

I. Introduction

"Hyperfine structure," like so much of the terminology of the quantum theory of matter, had its origin in atomic spectroscopy. "Structure" refers to the composite nature of spectral lines as revealed under high resolution. The adjective describes the fineness of resolution necessary to discern certain features of line structure. To illustrate, the *D* line emitted by a sodium vapor lamp (due to the $^2P \rightarrow {}^2S$ transition at approximately 25,000 cm^{-1}) is found under high resolution to be a doublet, split by 17 cm^{-1}. This feature is known as fine structure. Fine structure or multiplet structure in atoms is known to arise from magnetic interactions among internal angular momenta. For example, in sodium, the ^2P term is split by spin-orbit coupling into two levels, $^2P_{3/2}$ and $^2P_{1/2}$, giving two closely spaced emission lines. Under extremely high resolution, each component of the doublet appears as a complex pattern of narrow lines separated by intervals of the order of 0.05 cm^{-1}. This is hyperfine structure (hfs). The interactions leading to the latter splittings span the energy range 1–10,000 Mc/sec (1 $\text{cm}^{-1} \approx 30{,}000$ Mc/sec). Hyperfine structure, observed even before 1900 (Michelson, 1891; Fabry and Perot, 1897), was attributed by Pauli (1924) to interactions of atomic electrons with nuclear magnetic moments. A theoretical interpretation of the characteristic patterns of hfs intervals was developed by Back and Goudsmit (1927, 1928). This theory

---

* Present address: Department of Chemistry, University of Michigan, Ann Arbor, Michigan.

was closely analogous to the vector coupling model for multiplet structure (Goudsmit and Bacher, 1929). More rigorous quantum-mechanical treatments of hyperfine coupling were given in the next few years by Hargreaves (1929, 1930), Fermi (1930), and others (Breit, 1930, 1931a; Casimir, 1930; Goudsmit, 1931).

Higher nuclear moments (electric quadrupole, magnetic octupole, etc.) are other possible sources of hyperfine structure. In molecular systems there are, in addition, interactions with rotational magnetic moments and interactions, both direct and indirect, involving more than one magnetic nucleus. We shall not consider the latter effects in this paper, restricting our considerations to magnetic dipolar interactions of nuclear moments with electron spin and orbital moments.

Although the orientation of this review is theoretical, we should enumerate briefly the principal experimental techniques in the study of hyperfine interactions. The classical method is, of course, optical spectroscopy. Resolution of hyperfine structure in optical spectra[1] has been improved significantly by the use of interferometers, such as the Fabry–Perot etalon, in place of diffraction gratings. Splittings as small as 0.005 cm^{-1} ($\sim$100 Mc/sec) could theoretically be resolved by an interference spectrograph. In practice, however, atoms in the gas phase have linewidths determined by the Doppler effect which are typically of the order of 0.1 cm^{-1} for light atoms to 0.01 cm^{-1} for heavy atoms. Use of atomic beams, in which Doppler broadening is eliminated by observing atoms perpendicular to their line of flight, is required for measurements of splittings finer than 0.01 cm^{-1}. The most successful technique for the study of atomic hfs has been atomic beam magnetic resonance, developed principally by I. I. Rabi and co-workers.[2] Paramagnetic resonance of atoms in the gas phase[3] and in inert matrices (Jen *et al.*, 1956, 1958; Foner *et al.*, 1958, 1960) and of transition metal ions in solution and in crystals (Griffith, 1961, Chapter 12) has also yielded useful data on hyperfine interactions. It must, however, be recognized that external interactions (crystal fields; solvent, matrix, and pressure effects) usually influence the observed hyperfine structure. Finally, we mention the optical polarization spin-exchange technique (Brossel and

---

[1] For reviews of optical studies of hfs see White (1934), Chapter XVIII, and Kopfermann (1958), p.82ff.

[2] Techniques of atomic beam measurements and tabulated results are given in Kopfermann (1958) and in Ramsey (1955).

[3] Hydrogen, deuterium, oxygen, nitrogen, phosphorus, and iodine atoms have been studied by paramagnetic resonance: Brix (1952); Beringer and Heald (1954); Heald and Beringer (1954); Wittke and Dicke (1954, 1956); Dehmelt (1955); Bowers *et al.* (1957).

Kastler, 1949; Brossel *et al.*, 1950) for S state atoms which has yielded the most accurate hfs measurements to date for hydrogen, deuterium, nitrogen, and phosphorus atoms (Holloway and Novick, 1958; L. W. Anderson *et al.*, 1958, 1959, 1960a,b; Pipkin and Lambert, 1962; Lambert and Pipkin, 1962).

II. Relativistic Theory of the Electron[4]

In order to understand magnetic interactions from a fundamental point of view it is necessary to begin with a formulation of the quantum theory of the electron which accounts explicitly for its intrinsic (or spin) angular momentum and magnetic moment. The familiar nonrelativistic Schrödinger equation is not adequate in this respect. In the nonrelativistic theory, spin angular momentum and magnetic moment are introduced as auxiliary postulates. The resulting patchwork—the Pauli–Schrödinger theory—provides an adequate description for all but the finest features of atomic and molecular structure. The effects which concern us here are, however, of the latter category and we must begin with a more fundamental theory of the electron.

As soon as we consider magnetic interactions of any sort we are inextricably involved with relativity. To see this we recall that magnetic effects are represented by introduction into the Hamiltonian of terms of the type $(e/c)\mathbf{A}$, where $\mathbf{A}$ is the vector potential. In nonrelativistic mechanics $c \to \infty$ so that we could not have any magnetic effects in a rigorously consistent nonrelativistic theory. Conversely, if we admit magnetic effects we must make our theory fully consistent with special relativity. The nonrelativistic Schrödinger equation for a free particle

$$i\hbar \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m} \nabla^2 \psi \tag{1}$$

is not satisfactory from a relativistic point of view since space and time enter in a nonsymmetrical fashion—second derivatives wrt (with respect to) space variables but only a first derivative wrt time.

The Schrödinger equation is a quantum-mechanical transcription of the classical nonrelativistic energy–momentum relation

$$E = \frac{p^2}{2m} \tag{2}$$

---

[4] For more comprehensive treatments of the relativistic theory of the electron see Schiff (1955), Chapter XII; Dirac (1958), Chapter XI; Mandl (1957), Chapter X.

in which we make the formal substitution of differential operators

$$i\hbar\frac{\partial}{\partial t} \quad \text{for} \quad E$$

$$-i\hbar\nabla \quad \text{for} \quad \mathbf{p}$$

and allow the resulting operators to act on the wave function ψ. We can construct a covariant (relativistically correct) wave equation if we begin with the relativistic energy–momentum relation

$$E^2 = c^2p^2 + m^2c^4 \tag{3}$$

and make the same operator substitutions. The resulting relativistic wave equation

$$-\hbar^2\frac{\partial^2\psi}{\partial t^2} = \{-\hbar^2c^2\nabla^2 + m^2c^4\}\psi \tag{4}$$

was also given originally by Schrödinger (1926). Subsequently the identical equation was proposed independently by Gordon (1926), Klein (1927), and Fock (1926a,b). It is usually known as the Klein–Gordon equation. In (4) we have second derivatives in time to match the second derivatives in the space coordinates. It can be shown that (4) reduces to (1) in the nonrelativistic ($c \to \infty$) limit. Unfortunately, however, the Klein–Gordon equation, even though it is consistent with special relativity, does not seem to describe the electron correctly. It predicts the Sommerfeld fine-structure levels incorrectly but, more important, nowhere in the theory does spin appear. It is now believed that π mesons and K particles obey the Klein–Gordon equation. But, in any case, it is not the correct wave equation for an electron.

It was the brilliant idea of Dirac in 1928 to write a covariant wave equation containing *first* rather than second derivatives in the space and time variables. One could set up such an equation by the quantization procedure leading to (1) or (4) if there existed a linear relation between energy and momentum analogous to (3). We can, in fact, obtain a formal linearized version of (3) by introducing four higher algebraic quantities (hypercomplex numbers) α_x, α_y, α_z, and β which do not obey the commutative law of multiplication among themselves. We can then write

$$E = c\boldsymbol{\alpha}\cdot\mathbf{p} + \beta mc^2. \tag{5}$$

For our theory to be consistent we should recover (3) by taking the square of (5). This establishes a set of algebraic conditions on the α's and β:

$$\begin{aligned} E^2 = (c\boldsymbol{\alpha}\cdot\mathbf{p} + \beta mc^2)^2 &= c^2(\boldsymbol{\alpha}\cdot\mathbf{p})^2 + mc^3(\boldsymbol{\alpha}\cdot\mathbf{p})\beta + mc^3\beta(\boldsymbol{\alpha}\cdot\mathbf{p}) + \beta^2m^2c^4 \\ &= c^2p^2 + m^2c^4. \end{aligned} \tag{6}$$

Being careful to preserve the order of factors in multiplications, we deduce the following conditions for the consistency of (3) and (5). Denoting α_x, α_y, a_z, and β by α_1, α_2, α_3, and α_4, respectively:

$$\left.\begin{array}{ll} \alpha_k^2 = 1 & \\ \alpha_j\alpha_k + \alpha_k\alpha_j = 0 & k \neq j \end{array}\right\} j, k = 1, 2, 3, 4. \tag{7}$$

The α_k form a set of four self-inverse and anticommuting quantities. It is useful to have a matrix representation for them. By several algebraic theorems it may be shown that we require matrices no smaller than 4 × 4 (see, for example, Van der Waerden, 1932, Chapter 2). A particularly convenient representation is the following:[5]

$$\boldsymbol{\alpha} = \left(\begin{array}{c|c} 0 & \boldsymbol{\sigma} \\ \hline \boldsymbol{\sigma} & 0 \end{array}\right) \qquad \beta = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & -1 \end{array}\right) \tag{8}$$

where $\boldsymbol{\alpha}$ and β are written in partitioned form, each element representing a 2 × 2 matrix:

$$1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad 0 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \boldsymbol{\sigma} = \begin{pmatrix} \hat{k} & \hat{\imath} - i\hat{\jmath} \\ \hat{\imath} + i\hat{\jmath} & -\hat{k} \end{pmatrix}. \tag{9}$$

The components of $\boldsymbol{\sigma}$ are the well-known Pauli spin matrices

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{9a}$$

Quantizing the energy-momentum relation (5) we obtain the Dirac equation for a free particle

$$i\hbar \frac{\partial \Psi}{\partial t} = \{-i\hbar c\boldsymbol{\alpha}\cdot\nabla + \beta mc^2\}\Psi \tag{10}$$

which has the desired linearity in space and time derivatives. Because of the dimensionality of the α's and β, Ψ must be interpreted not as a scalar function of $\mathbf{r}$ and t but as a four-component column vector

$$\Psi = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} \tag{11}$$

---

[5] One should not attempt to interpret the classical relation (5) using this matrix representation.

known as a "spinor," where each X is an ordinary scalar function of **r** and t. Two-component spinors occur in the Pauli–Schrödinger theory, for we could write the wave functions of an electron in spin states $m_s = \pm\frac{1}{2}$ as

$$\Psi = \mathrm{X}(\mathbf{r}, t)\begin{pmatrix}1\\0\end{pmatrix}$$

and (12)

$$\Psi = \mathrm{X}(\mathbf{r}, t)\begin{pmatrix}0\\1\end{pmatrix},$$

respectively. In general, for a superposition of spin eigenstates, we would have

$$\Psi = \begin{pmatrix}\mathrm{X}_1(\mathbf{r}, t)\\ \mathrm{X}_2(\mathbf{r}, t)\end{pmatrix} \tag{13}$$

in complete analogy to (11). For a spinor wave function the probability density is given by

$$\rho(\mathbf{r}, t) = \Psi^+\Psi = \mathrm{X}_1^*\mathrm{X}_1 + \mathrm{X}_2^*\mathrm{X}_2 + \cdots \tag{14}$$

where Ψ^+ is the hermitian adjoint (transposed conjugate) of Ψ. In the Pauli–Schrödinger theory the components X_1 and X_2 of (13) determine the relative weights of the two internal states of the electron—spin up and spin down. Interpreting (11) analogously, we conclude that a Dirac particle has four internal states. It can be shown that two of these are states in which the particle has the character of an electron and the other two that of a positron. For a free particle it is always possible to choose a representation in which only one of the components of the spinor (11) is different from zero. For a particle in a field, however, this is not the case and we may have a Dirac particle partaking of the characteristics of both electron and positron. This is perhaps somewhat akin to the concept of resonance in chemistry. In the nonrelativistic limit—i.e., particle velocities $v \ll c$ and energies $E \ll mc^2$—the positron components of the spinor are much smaller than the electron components—by a factor of order v/c—and may be neglected in our work.

So far we have written the Dirac equation for a free particle. For an electron in a field we make the formal substitutions in both classical and quantum theory

$$E \rightarrow E + e\phi, \qquad \mathbf{p} \rightarrow \boldsymbol{\pi} = \mathbf{p} + \frac{e}{c}\mathbf{A} \tag{15}$$

where $\phi(\mathbf{r})$ is the scalar (electrostatic) potential, $\mathbf{A}(\mathbf{r})$ the vector potential, and $-e$ the electronic charge. We shall consider the interaction of an electron with the electric and magnetic fields of a nucleus. We have then for the Dirac Equation in a field $(\phi, \mathbf{A})$

$$i\hbar\frac{\partial\Psi}{\partial t} = \{c\boldsymbol{\alpha}\cdot\boldsymbol{\pi} + \beta mc^2 - e\phi\}\Psi. \tag{16}$$

For a stationary (time-independent) solution

$$i\hbar\frac{\partial\Psi}{\partial t} = \mathcal{H}\Psi = E_R\Psi. \tag{17}$$

We write E_R for the energy eigenvalue to emphasize that it is a relativistic energy, including the rest energy mc^2. Defining the nonrelativistic energy E by

$$E = E_R - mc^2 \tag{18}$$

we have then, for stationary states,

$$\{c\boldsymbol{\alpha}\cdot\boldsymbol{\pi} + (\beta - 1)mc^2 - e\phi\}\Psi = E\Psi. \tag{19}$$

Since the α's and β are 4×4 matrices, this matrix equation is an abbreviation for four simultaneous scalar equations for the four components of Ψ. It is convenient to rewrite (19) as two simultaneous equations involving 2×2 matrices. Accordingly we represent Ψ by

$$\Psi = \begin{pmatrix}\Psi_1\\ \hline \Psi_2\end{pmatrix} \quad \text{where} \quad \Psi_1 = \begin{pmatrix}X_1\\ X_2\end{pmatrix} \quad \text{and} \quad \Psi_2 = \begin{pmatrix}X_3\\ X_4\end{pmatrix}. \tag{20}$$

Recalling the definitions of α and β we have for (19) in partitioned matrix form:

$$\left[\left(\begin{array}{c|c}0 & c\boldsymbol{\sigma}\cdot\boldsymbol{\pi}\\ \hline c\boldsymbol{\sigma}\cdot\boldsymbol{\pi} & 0\end{array}\right) + \left(\begin{array}{c|c}0 & 0\\ \hline 0 & -2mc^2\end{array}\right) - \left(\begin{array}{c|c}e\phi & 0\\ \hline 0 & e\phi\end{array}\right)\right]\begin{pmatrix}\Psi_1\\ \hline \Psi_2\end{pmatrix} = \left(\begin{array}{c|c}E & 0\\ \hline 0 & E\end{array}\right)\begin{pmatrix}\Psi_1\\ \hline \Psi_2\end{pmatrix} \tag{21}$$

which may be written as coupled equations for Ψ_1, and Ψ_2:

$$\begin{aligned}(E + e\phi)\Psi_1 - c\boldsymbol{\sigma}\cdot\boldsymbol{\pi}\Psi_2 &= 0\\ (E + e\phi + 2mc^2)\Psi_2 - c\boldsymbol{\sigma}\cdot\boldsymbol{\pi}\Psi_1 &= 0.\end{aligned} \tag{22}$$

In the nonrelativistic limit E and $e\phi$ are of the order $mv^2(\ll mc^2)$ and π is of the order mv. Thus components of Ψ_2 (the positron components) are evidently smaller than those of Ψ_1 by a factor of order v/c.[6] The effect of Ψ_2 on energy quantities is smaller still since the latter involve squares of the spinor components. Ψ_2 may be eliminated between Eqs. (22) to yield an equation for Ψ_1 alone. For

$$\Psi_2 = \frac{1}{2mc}\,k(\mathbf{r})\boldsymbol{\sigma}\cdot\boldsymbol{\pi}\Psi_1 \tag{23}$$

having defined the function

$$k(\mathbf{r}) = \left[1 + \frac{E + e\phi(\mathbf{r})}{2mc^2}\right]^{-1}. \tag{24}$$

Ψ_1 then satisfies the wave equation

$$\left\{E + c\phi - \frac{1}{2m}\,\boldsymbol{\sigma}\cdot\boldsymbol{\pi}k(\mathbf{r})\boldsymbol{\sigma}\cdot\boldsymbol{\pi}\right\}\Psi_1 = 0 \tag{25}$$

which we may write

$$\mathscr{H}_{\text{eff}}\Psi_1 = E\Psi_1. \tag{25a}$$

We shall now forget entirely about Ψ_2 and regard (25) as the fundamental wave equation in a two-component theory. To a degree of approximation obtained by neglecting contributions second order and higher in the fine-structure constant α ($\alpha = e^2/\hbar c \approx 1/137.02 \approx v/c$ for an atomic electron), Ψ_1 is equivalent to the Pauli-Schrödinger wave function. We have improved upon the original Pauli–Schrödinger formulation in that (*i*) we have arrived at it by a consistent, unified approach and (*ii*) $\mathscr{H}_{\text{eff}}$ contains certain terms missing from the original theory, in which $k(\mathbf{r})$ is simply replaced by 1. The missing terms are, in fact, vital to an understanding of hyperfine structure.

Substituting in the effective Hamiltonian the definition (15) of $\boldsymbol{\pi}$ we find

$$\mathscr{H}_{\text{eff}} = \mathscr{H}^{(0)} + \mathscr{H}^{(1)} + \mathscr{H}^{(2)} \tag{26}$$

where

$$\mathscr{H}^{(0)} = -e\phi + \frac{1}{2m}\,\boldsymbol{\sigma}\cdot\mathbf{p}k(\mathbf{r})\boldsymbol{\sigma}\cdot\mathbf{p}, \tag{27}$$

$$\mathscr{H}^{(1)} = \frac{e}{2mc}\,\{\boldsymbol{\sigma}\cdot\mathbf{p}k(\mathbf{r})\boldsymbol{\sigma}\cdot\mathbf{A} + \boldsymbol{\sigma}\cdot\mathbf{A}k(\mathbf{r})\boldsymbol{\sigma}\cdot\mathbf{p}\}, \tag{28}$$

$$\mathscr{H}^{(2)} = \frac{e^2}{2mc^2}\,\boldsymbol{\sigma}\cdot\mathbf{A}k(\mathbf{r})\boldsymbol{\sigma}\cdot\mathbf{A}. \tag{29}$$

---

[6] $\Psi_2 \sim (v/c)\Psi_1$ when the Hamiltonian is written as $\mathscr{H} = c\boldsymbol{\alpha}\cdot\boldsymbol{\pi} + \beta mc^2 - e\phi$. The alternative choice, $\mathscr{H} = -c\boldsymbol{\alpha}\cdot\boldsymbol{\pi} - \beta mc^2 - e\phi$, which occurs frequently, reverses the small and large components.

The superscript gives the order of the operator *wnt* the fine-structure constant. $\mathscr{H}^{(0)}$ contains the ordinary potential and kinetic energies of the electron (with relativistic corrections) plus the spin-orbit coupling. In the absence of a magnetic field ($\mathbf{A} = 0$), $\mathscr{H}_{\text{eff}}$ reduces the $\mathscr{H}^{(0)}$. In a central field [$\mathbf{A} = 0$ and $\phi = \phi(r)$, hence $k = k(r)$] we might expect in analogy with nonrelativistic theory that the electron's angular momentum is a constant of the motion, i.e., that it commutes with $\mathscr{H}^{(0)}$. We find however that

$$[\mathbf{l}, \mathscr{H}^{(0)}] = [-\tfrac{1}{2}\boldsymbol{\sigma}, \mathscr{H}^{(0)}] \neq 0. \tag{30}$$

The electron's orbital angular momentum, $\mathbf{l} = \hbar^{-1}\mathbf{r} \times \mathbf{p} = -i\mathbf{r} \times \nabla$, is not a conserved quantity. However, we see that the operator $\mathbf{j} = \mathbf{l} + \frac{1}{2}\boldsymbol{\sigma}$ does commute with the Hamiltonian. We are thus led to the conclusion that the electron has an intrinsic angular momentum

$$\mathbf{s} = \tfrac{1}{2}\boldsymbol{\sigma} \quad \text{(in units of } \hbar) \tag{31}$$

and that only the total angular momentum

$$\mathbf{j} = \mathbf{l} + \mathbf{s} \tag{32}$$

is rigorously a constant of the motion. This result also follows in the original four-component Dirac theory, where s is given instead by

$$\mathbf{s} = \frac{1}{2}\left(\begin{array}{c|c} \boldsymbol{\sigma} & 0 \\ \hline 0 & \boldsymbol{\sigma} \end{array}\right). \tag{33}$$

III. The Hamiltonian for Hyperfine Interactions[7]

This is contained in the first-order term $\mathscr{H}^{(1)}$. We now substitute explicit forms for the electromagnetic potentials ϕ and $\mathbf{A}$. For a nucleus of charge Ze and magnetic moment $\boldsymbol{\mu}$,

$$\begin{aligned} \phi &= Zer^{-1} \\ \mathbf{A} &= \operatorname{curl}\frac{\boldsymbol{\mu}}{r} = r^{-3}\boldsymbol{\mu} \times \mathbf{r} \end{aligned} \tag{34}$$

just as for a classical point charge and point magnetic dipole. Even though we neglect the quantum nature of the nucleus, regarding it merely as a field

---

[7] Alternative derivations may be found in Bethe (1933), p.385; Griffith (1961), Chapters 5 and 12; Blinder (1960a). A somewhat different approach is used by Frosch and Foley (1952).

source, we will treat the nuclear spin **I** as a quantum-mechanical operator.[8] $\boldsymbol{\mu}$ and **I** are related by

$$\boldsymbol{\mu} = g_I \mu_N \mathbf{I} \tag{35}$$

where $\mu_N = e\hbar/2Mc$ (the nuclear magneton) and g_I is the nuclear g factor referred to μ_N. In alternative notation we have

$$\boldsymbol{\mu} = \hbar \gamma_I \mathbf{I} \tag{35a}$$

where γ_I is called the gyromagnetic (or magnetogyric) ratio. Since **p** is the operator $-i\hbar\nabla$, we have

$$\mathbf{p}k(r) = k(r)\mathbf{p} - i\hbar \frac{\mathbf{r}}{r}\frac{dk}{dr}. \tag{36}$$

We require also the identity due to Dirac

$$\boldsymbol{\sigma}\cdot\mathbf{a}\,\boldsymbol{\sigma}\cdot\mathbf{b} = \mathbf{a}\cdot\mathbf{b} + i\boldsymbol{\sigma}\cdot\mathbf{a}\times\mathbf{b} \tag{37}$$

for two arbitrary vectors **a** and **b** commuting with $\boldsymbol{\sigma}$. Applying (36) and (37) to (28) we find

$$\mathscr{H}^{(1)} = \frac{e}{2mc}\left\{k(r)[\mathbf{p}\cdot\mathbf{A} + \mathbf{A}\cdot\mathbf{p} + i\boldsymbol{\sigma}\cdot(\mathbf{p}\times\mathbf{A} + \mathbf{A}\times\mathbf{p})] \right.$$
$$\left. + \frac{dk}{dr}\left[-\frac{i\hbar}{r}\mathbf{r}\cdot\mathbf{A} + \frac{\hbar}{r}\boldsymbol{\sigma}\cdot\mathbf{r}\times\mathbf{A}\right]\right\}. \tag{38}$$

Substituting the explicit form for **A** from (34) we obtain, after considerable vector manipulation,[9]

$$\mathscr{H}^{(1)} = -gg_I\mu_0\mu_N\left\{k(r)[r^{-3}\mathbf{I}\cdot(\mathbf{l}-\mathbf{s}) + 3r^{-5}\mathbf{I}\cdot\mathbf{r}\,\mathbf{s}\cdot\mathbf{r}] \right.$$
$$\left. + \frac{dk}{dr}[r^{-2}\mathbf{I}\cdot\mathbf{s} - r^{-4}\mathbf{I}\cdot\mathbf{r}\,\mathbf{s}\cdot\mathbf{r}]\right\}. \tag{39}$$

We have employed the definitions

$$\hbar\mathbf{l} = \mathbf{r}\times\mathbf{p}$$
$$\mathbf{s} = \tfrac{1}{2}\boldsymbol{\sigma} \tag{40}$$
$$\mu_0 = e\hbar/2mc \quad \text{(the Bohr magneton)}$$

and written g in place of a factor -2 for the electron's g factor. Equation (39) contains, for a one-electron system, to first-order accuracy in the

---

[8] A more correct approach, employing a two-particle relativistic equation, is given by Arnowitt (1953). See also Newcomb and Salpeter (1955).

[9] A helpful list of vector identities is given in Ramsey (1955), Appendix B.

fine-structure constant, all the interactions coupling a nuclear moment to electronic spin and orbital moments. $\mathscr{H}^{(2)}$ does not contribute any terms of this type since from (29) and (37)

$$\mathscr{H}^{(2)} = \frac{e^2}{2mc^2} k(r) A^2 \tag{41}$$

independent of the electron's spin or orbital momenta. For the vector potential of a magnetic dipole (34), (41) when evaluated for an *s*-electron leads to a divergent integral. This is just the infinite self-energy of a point magnetic dipole (the same type of difficulty arises also for point charges). The trouble disappears if a fully covariant two-particle theory is employed but, in any case, it does not affect our considerations of $\mathscr{H}^{(1)}$.

Certain aspects of atomic and molecular structure suggest further approximations to (39). Let us consider first the properties of $k(r)$ and its derivative. From (24) and (34) we have

$$\begin{aligned} k(r) &= \left[1 + \frac{E + e\phi(r)}{2mc^2}\right]^{-1} \approx \left(1 + \frac{r_0}{r}\right)^{-1} \\ k'(r) &= \left(1 + \frac{r_0}{r}\right)^{-2} \frac{r_0}{r^2}. \end{aligned} \tag{42}$$

These functions are sketched in Fig. 1. We have defined the parameter $r_0 = Ze^2/2mc^2 = 1.4089Z \times 10^{-13}$ cm, which is

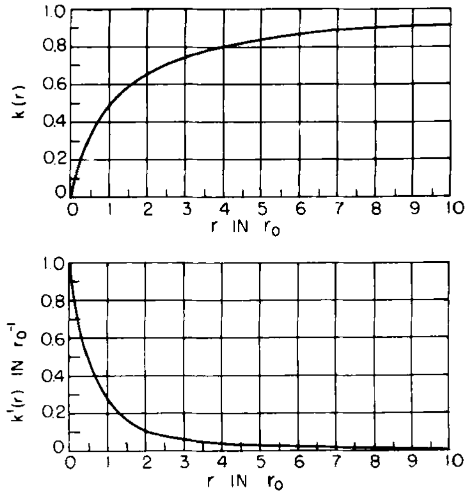


Fig. 1. Graphs of the functions $k(r)$ and $k'(r)$.

of the order of a nuclear diameter. Over a region of this dimension—or even 100 times larger—an electronic wave function changes minutely and may, to good approximation, be regarded as constant. $k(r)$ differs significantly from unity and $k'(r)$ from zero only in the immediate neighborhood of the nucleus. We note that unless the electron is effectively excluded from this region, as in the case of a non-s atomic orbital, the function $k(r)$ cannot be expanded in inverse powers of mc^2 with the expectation that successive terms will converge, for $e\phi$ becomes comparable to mc^2 when $r \approx r_0$.

Since $k(r)$ approaches zero linearly in r, the expectation value of the operator

$$k(r)[r^{-3}\mathbf{I}\cdot(\mathbf{l}-\mathbf{s}) + 3r^{-5}\mathbf{I}\cdot\mathbf{r}\,\mathbf{s}\cdot\mathbf{r}] \tag{43}$$

will have zero contribution from the neighborhood of $r = 0$, provided only that the wave function is finite at $r = 0$. Since, as we have seen, $k(r) \approx 1$ elsewhere, we may make the approximate replacement for (43) of

$$[r^{-3}\mathbf{I}\cdot(\mathbf{l}-\mathbf{s}) + 3r^{-5}\mathbf{I}\cdot\mathbf{r}\,\mathbf{s}\cdot\mathbf{r}]_{r>0} \tag{44}$$

indicating that only the principal part of the integral without $k(r)$ is to be evaluated.[10] Note that this operator averages to zero for an s atomic orbital.

The second bracket of (39) may likewise be simplified. We note first that $k'(r)$ is effectively a delta function for

$$k'(r) \approx 0 \qquad r \not\approx 0$$

and

$$\int_0^\infty dr\, k'(r) = k(\infty) - k(0) = 1. \tag{45}$$

---

[10] This may be done more precisely as follows. The radial integral in a matrix element of (43) will be of the form

$$\int_0^\infty dr\, r^2 r^{-3} k(r) R^2(r) \dots$$

where $R(r)$ is the radial part of the wave function. From (42) we have

$$r^{-1}k(r) = (r + r_0)^{-1}.$$

Thus the above integral reduces to

$$\int_0^\infty dr (r + r_0)^{-1} R^2(r) \dots .$$

Changing the variable of integration this becomes

$$\int_{r_0}^\infty dr\, r^{-1} R^2(r - r_0).$$

But $R(r - r_0) \approx R(r)$ since atomic and molecular wave functions are practically constant over a region of dimension r_0.

In terms of the three-dimension delta function $\delta(\mathbf{r})$ we have

$$k'(r) \approx \delta(r) = 4\pi r^2 \delta(\mathbf{r}). \tag{46}$$

When the bracketed expression is averaged over all angles (e.g., in expectation values) we find

$$[r^{-2}\mathbf{I}\cdot\mathbf{s} - r^{-4}\mathbf{I}\cdot\mathbf{r}\,\mathbf{s}\cdot\mathbf{r}]_{\mathrm{av}} = \tfrac{2}{3}r^{-2}\mathbf{I}\cdot\mathbf{s}. \tag{47}$$

Combining (46) with (47) we are led to the extremely useful approximation

$$k'(r)[r^{-2}\mathbf{I}\cdot\mathbf{s} - r^{-4}\mathbf{I}\cdot\mathbf{r}\,\mathbf{s}\cdot\mathbf{r}] \approx \frac{8\pi}{3}\,\delta(\mathbf{r})\mathbf{I}\cdot\mathbf{s}. \tag{48}$$

This is the well-known Fermi contact interaction (Fermi, 1930; Fermi and Segre, 1933a,b), so-called because it is effective only at extremely short range.[11] Since only s atomic orbitals have nonzero densities at the nucleus, only an orbital containing some s character can contribute to the Fermi term, the contribution being then proportional to $|\psi(0)|^2$.

To summarize, we have reduced the hyperfine interaction operator to the approximate form

$$\mathscr{H}^{(1)} \approx -gg_I\mu_0\mu_N\left\{[r^{-3}\mathbf{I}\cdot(\mathbf{l} - \mathbf{s}) + 3r^{-5}\mathbf{I}\cdot\mathbf{r}\,\mathbf{s}\cdot\mathbf{r}]_{r>0} + \frac{8\pi}{3}\,\delta(\mathbf{r})\mathbf{I}\cdot\mathbf{s}\right\}. \tag{49}$$

For future reference we write $\mathscr{H}^{(1)}$ as the sum

$$\mathscr{H}^{(1)} = \mathscr{H}_L^{(1)} + \mathscr{H}_D^{(1)} + \mathscr{H}_C^{(1)} \tag{49a}$$

with

$$\begin{aligned} \mathscr{H}_L^{(1)} &= -gg_I\mu_0\mu_N r^{-3}\mathbf{I}\cdot\mathbf{l}, \\ \mathscr{H}_D^{(1)} &= -gg_I\mu_0\mu_N[3r^{-5}\mathbf{I}\cdot\mathbf{r}\,\mathbf{s}\cdot\mathbf{r} - r^{-3}\mathbf{I}\cdot\mathbf{s}]_{r>0}, \\ \mathscr{H}_C^{(1)} &= -gg_I\mu_0\mu_N \frac{8\pi}{3}\,\delta(\mathbf{r})\mathbf{I}\cdot\mathbf{s}. \end{aligned} \tag{49b}$$

We shall denote these as the orbital, dipolar, and contact operators, respectively.

We have been implicitly assuming that expectation values of $\mathscr{H}^{(1)}$ are to be evaluated using solutions of the nonrelativistic Schrödinger equation. This is equivalent to neglect of all contributions of order $\alpha^2 Z^2$ smaller than those considered—an adequate approximation except when we are dealing with inner shells of heavy atoms ($Z \gg 1$). In the latter case, a more complete

---

[11] The large contribution of unpaired s electrons to hfs was noted by Meggers and Burns (1927).

relativistic treatment is required. The small components of the Dirac spinor (which, as we have seen, give terms of order $\alpha^2 Z^2$) can no longer be thrown away. There is an added complication when solutions of the Dirac equation are used, for wave functions with $j = \frac{1}{2}$ are singular at $r = 0$, so that the approximations leading to (49) are invalid. The singularities are sufficiently weak however, being of the form $r^{[(1-\alpha^2Z^2)^{\frac{1}{2}}-1]} \approx r^{-\frac{1}{2}\alpha^2Z^2}$, so that all integrals are convergent, provided that the operator (49) is replaced by its more rigorous analog (39). Relativistic corrections to hyperfine structure are discussed by Breit (1930, 1931b), Racah (1931, 1932), and others (Goudsmit, 1933; Fermi and Segre, 1933a,b).[12] There are additional corrections in heavy atoms for finite nuclear size (Rosenthal and Breit, 1932; Crawford and Schawlow, 1949; Bohr and Weisskopf, 1950; Sessler and Mills, 1958).

Our approximation procedure also depends on the inequality

$$\frac{1}{|\psi|}\left|\frac{\partial\psi}{\partial r}\right| r_0 \ll 1, \tag{50}$$

which is indeed true for electronic distributions in all atoms and molecules. Finally, we note that the reduction leading from (39) to (49) would take a different form if the scalar potential $\phi(\mathbf{r})$ were other than Coulombic, notwithstanding the fact that the potential does not enter explicitly in (49). In particular, for ϕ proportional to $1/r^n$ with $n > 1$ the delta function term would not appear.

It is interesting that (49) may be derived directly from classical theory without use of relativistic quantum mechanics (Ferrell, 1960; Milford, 1960; Rado, 1962). We first evaluate the magnetic induction $\mathbf{B}$, due to a point magnetic dipole $\boldsymbol{\mu}$:

$$\mathbf{B} = \text{curl}\,\mathbf{A} = \text{curl curl}\,\frac{\boldsymbol{\mu}}{r} = \text{grad div}\,\frac{\boldsymbol{\mu}}{r} - \nabla^2\frac{\boldsymbol{\mu}}{r}. \tag{51}$$

But

$$\text{div}\,\frac{\boldsymbol{\mu}}{r} = \boldsymbol{\mu}\cdot\text{grad}\,\frac{1}{r} = -\frac{\boldsymbol{\mu}\cdot\mathbf{r}}{r^3} \tag{52}$$

and

$$\frac{\partial}{\partial x}\left(-\frac{\boldsymbol{\mu}\cdot\mathbf{r}}{r^3}\right) = -\frac{\mu_x}{r^3} + \frac{3(\boldsymbol{\mu}\cdot\mathbf{r})x}{r^5}. \tag{53}$$

Thus

$$\text{grad div}\,\frac{\boldsymbol{\mu}}{r} = -\frac{\boldsymbol{\mu}}{r^3} + \frac{3(\boldsymbol{\mu}\cdot\mathbf{r})\mathbf{r}}{r^5}. \tag{54}$$

---

[12] Relativistic correction factors are tabulated in Kopfermann (1958), Sect. 26.

Also

$$\nabla^2 \frac{\boldsymbol{\mu}}{r} = \boldsymbol{\mu}\, \nabla^2 \frac{1}{r} = -4\pi\boldsymbol{\mu}\delta(\mathbf{r}). \tag{55}$$

$\nabla^2(1/r)$ may be replaced by $-4\pi\delta(\mathbf{r})$ since

$$\nabla^2 \frac{1}{r} = 0, \qquad r \neq 0,$$

and

$$\int d\tau\, \nabla^2 \frac{1}{r} = \int \mathbf{dS}\cdot\operatorname{grad} \frac{1}{r} = 4\pi r^2\left(-\frac{r}{r^3}\right) = -4\pi, \tag{56}$$

the last integral being evaluated by means of the divergence theorem. We have then that

$$\mathbf{B} = \left[-\frac{\boldsymbol{\mu}}{r^3} + \frac{3(\boldsymbol{\mu}\cdot\mathbf{r})\mathbf{r}}{r^5}\right] + 4\pi\boldsymbol{\mu}\delta(\mathbf{r}). \tag{57}$$

This is consistent with the relation between magnetic induction **B** and magnetic field **H**

$$\mathbf{B} = \mathbf{H} + 4\pi\mathbf{M} \tag{58}$$

where **M** is the magnetization or magnetic moment per unit volume. The bracketed part of (57) is the familiar expression for the field of a dipole, whereas the magnetic moment per unit volume for a point dipole $\boldsymbol{\mu}$ is obviously $\boldsymbol{\mu}\delta(\mathbf{r})$. The energy of interaction between two point dipoles $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ is then given by

$$E = -\boldsymbol{\mu}_1\cdot\mathbf{B}_2 = \left[\frac{\boldsymbol{\mu}_1\cdot\boldsymbol{\mu}_2}{r^3} - \frac{3\boldsymbol{\mu}_1\cdot\mathbf{r}\boldsymbol{\mu}_2\cdot\mathbf{r}}{r^5}\right] - 4\pi\delta(\mathbf{r})\boldsymbol{\mu}_1\cdot\boldsymbol{\mu}_2. \tag{59}$$

But there is another delta function hidden in the bracket! To see this we note that

$$\begin{aligned}\frac{\boldsymbol{\mu}_1\cdot\boldsymbol{\mu}_2}{r^3} - \frac{3\boldsymbol{\mu}_1\cdot\mathbf{r}\boldsymbol{\mu}_2\cdot\mathbf{r}}{r^5} &= -\boldsymbol{\mu}_1\cdot\left\{\operatorname{grad}\operatorname{grad}\frac{1}{r}\right\}\cdot\boldsymbol{\mu}_2 \\ &= -\sum_{i,j=1}^{3} \boldsymbol{\mu}_{1i}\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}\frac{1}{r}\,\mu_{2j}.\end{aligned} \tag{60}$$

As we take the limit of this expression as $r \to 0$, only the isotropic part contributes. Averaging over all angles we find that

$$\begin{aligned}&\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}\frac{1}{r} \to 0 \qquad i \neq j \\ &\frac{\partial^2}{\partial x_i^2}\frac{1}{r} \to \frac{1}{3}\nabla^2\frac{1}{r} = -\frac{4\pi}{3}\delta(\mathbf{r}).\end{aligned} \tag{61}$$

Thus we have

$$\lim_{r\to 0}\left[\frac{\boldsymbol{\mu}_1\cdot\boldsymbol{\mu}_2}{r^3}-\frac{3\boldsymbol{\mu}_1\cdot\mathbf{r}\boldsymbol{\mu}_2\cdot\mathbf{r}}{r^5}\right]=-\frac{1}{3}\operatorname{tr}\left\{\operatorname{grad}\operatorname{grad}\frac{1}{r}\right\}\boldsymbol{\mu}_1\cdot\boldsymbol{\mu}_2$$

$$=\frac{4\pi}{3}\,\delta(\mathbf{r})\boldsymbol{\mu}_1\cdot\boldsymbol{\mu}_2. \tag{62}$$

Combining (62) with (59) we obtain

$$E=\left[\frac{\boldsymbol{\mu}_1\cdot\boldsymbol{\mu}_2}{r^3}-\frac{3\boldsymbol{\mu}_1\cdot\mathbf{r}\boldsymbol{\mu}_2\cdot\mathbf{r}}{r^5}\right]_{r>0}-\frac{8\pi}{3}\,\delta(\mathbf{r})\boldsymbol{\mu}_1\cdot\boldsymbol{\mu}_2. \tag{63}$$

Replacing $\boldsymbol{\mu}_2$ by $g_I\mu_N\mathbf{I}$ and $\boldsymbol{\mu}_1$ by $g\mu_0\mathbf{s}$, we recover all but the orbital part of (49). The orbital contribution is easily derived. In classical electrodynamics the interaction energy of an electron with an external magnetic field is given by

$$E=\frac{e}{c}\,\mathbf{v}\cdot\mathbf{A}. \tag{64}$$

Substituting (34) for **A** we find

$$E=\frac{e}{c}\frac{\mathbf{p}}{m}\cdot\frac{\boldsymbol{\mu}\times\mathbf{r}}{r^3}=\frac{e\hbar}{mc}\,r^{-3}\boldsymbol{\mu}\cdot\mathbf{l}=-gg_I\mu_0\mu_N r^{-3}\,\mathbf{I}\cdot\mathbf{l} \tag{65}$$

Since this term contributes nothing at $r=0$, we may insert it in the bracket of (63), thus completing the classical derivation of (49).

IV. Hyperfine Structure of One-Electron Atoms

The hyperfine interactions contained in $\mathscr{H}^{(1)}$ represent a rather small perturbation on an atomic system. Hence we shall be concerned only with first-order perturbation corrections, which are given by diagonal matrix elements of $\mathscr{H}^{(1)}$. The operators $\mathscr{H}^{(0)}$, k, j^2, j_z, I^2, and I_z constitute a complete compatible set of observables for the unperturbed atom. Thus, unperturbed atomic states may be specified by the set of quantum numbers n, k, j, m_j, I, and m_I. k is an operator, related to the orbital angular momentum which distinguishes between degenerate states of the same n and j. For example, the two states of the hydrogen atom known as $2S_{1/2}$ and $2P_{1/2}$ are degenerate in the Dirac theory (neglecting the Lamb shift) and differ only in their values of k. For a one-electron atom s^2 and l^2 (though not s_z or l_z) are also conserved quantities. s^2 commutes rigorously with the Hamiltonian

$$\mathscr{H}=-e\phi+c\boldsymbol{\alpha}\cdot\boldsymbol{\pi}+\beta mc^2 \tag{66}$$

in the four-component theory as well as with $\mathscr{H}_{\text{eff}}$ in the two-component theory. l^2 on the other hand does not rigorously commute with (66) even for a central field [$\phi = \phi(r)$ and $\pi = \mathbf{p}$]. The noncommuting part is, however, rather small and is evidently discarded in the reduction to a two-component theory, for l^2 is found to commute with $\mathscr{H}^{(0)}$. Thus the $2S_{1/2}$ and $2P_{1/2}$ states of hydrogen are, to very good approximation, characterized by expectation values $\langle l^2 \rangle$ of 0 and 1 (1 + 1), respectively, and we may use l in place of k to label atomic states. In the one-electron case we have j and l always related by either

$$j = l \pm \tfrac{1}{2}.$$

It should be carefully noted that the above properties of s^2 and l^2 do not generalize for the many-electron case.

Since $\mathscr{H}^{(1)}$ contains three types of angular momentum operators, some results from the theory of angular momentum greatly facilitate our analysis. We recall the defining commutation relations for the components of an angular momentum operator

$$[j_x, j_y] = ij_z \quad \text{et cyc.} \tag{67}$$

These may be abbreviated by the single dyadic relation

$$[\mathbf{j}, \mathbf{j}] = i\mathscr{I} \times \mathbf{j} \tag{68}$$

where $\mathscr{I}$ is the unit dyadic, $\mathscr{I} = \hat{\imath}\hat{\imath} + \hat{\jmath}\hat{\jmath} + \hat{k}\hat{k}$. Operators $\mathbf{T}$ obeying the closely related commutation relations

$$[\mathbf{j}, \mathbf{T}] = i\mathscr{I} \times \mathbf{T} \tag{69a}$$

or

$$\left.\begin{aligned} [j_x, T_y] &= iT_z \\ [j_z, T_z] &= 0 \end{aligned}\right\} \text{et cyc.} \tag{69b}$$

with respect to an angular momentum $\mathbf{j}$ are frequently encountered. An operator satisfying (69) is said to be "of type T with respect to the angular momentum $\mathbf{j}$." A second class of operators is characterized by complete commutativity with the components of $\mathbf{j}$:

$$[C, \mathbf{j}] = 0. \tag{70}$$

We shall say that such an operator is "of type C wrt $\mathbf{j}$." C may be a scalar, vector, or tensor operator. For example, with respect to the orbital angular momentum $\mathbf{l}$, it may be verified that $\mathbf{r}$, $\mathbf{p}$, and $\mathbf{l}$ itself are of type T, $\mathbf{s}$ and $\mathbf{r}\cdot\mathbf{p}$ are of type C, whereas $\mathbf{j}$ and $\mathbf{s}\cdot\mathbf{l}$ are of neither class. A geometrical interpretation of these operator types can be made if we recall that $\mathbf{l}$ is a multiple of the rotation operator in three-dimensional space.[13] Any axial or polar vector lying entirely within this space is transformed under rotation in a manner consistent with the commutation relations (69) and is hence of type T wrt $\mathbf{l}$. On the other hand, any operator completely independent of this space, such as $\mathbf{s}$, or any scalar invariant, such as $\mathbf{r}\cdot\mathbf{p}$, is not affected by the rotation operator and hence belongs to the class C wrt $\mathbf{l}$. The operator $\mathbf{j}$ is related to rotation in a more abstract vector space, but an analogous geometrical interpretation is possible. The following set of rules for "building up" new operators of type T or C is readily verified:[14]

(1) if $\mathbf{T}$ satisfies (69) wrt $\mathbf{j}_1$ and commutes with $\mathbf{j}_2$ then $\mathbf{T}$ also satisfies (69) wrt $\mathbf{j}_1 + \mathbf{j}_2$;
(2) if $\mathbf{T}_1$ and $\mathbf{T}_2$ both satisfy (69), the $\mathbf{T}_1 + \mathbf{T}_2$ and $\mathbf{T}_1 \times \mathbf{T}_2$ are of type T wrt $\mathbf{j}$, whereas $\mathbf{T}_1 \cdot \mathbf{T}_2$ is of type C;
(3) if $\mathbf{T}$ satisfies (69) and A commutes with $\mathbf{j}$ then $A\mathbf{T}$ is of type T.

For example, with respect to $\mathbf{j}$, the vectors $\mathbf{r}$, $\mathbf{p}$, $\mathbf{l}$, $\mathbf{s}$, $\mathbf{j}$, $\mathbf{l}\times\mathbf{s}$ and $r^{-3}\mathbf{l}$ are of type T, $\mathbf{l}\cdot\mathbf{s}$, l^2, s^2 and $\mathbf{I}$ are of type C, while $\mathbf{I}\cdot\mathbf{s}$, $\mathbf{I}\cdot\mathbf{l}$ and $\mathbf{I}\cdot\mathbf{j}$ belong to neither class.

The evaluation of matrix elements of type T operators is discussed in detail in Condon and Shortley (1935). A particularly useful generalization is the "replacement theorem" (Griffith, 1961, p.38), which states that the m dependence of any type T operator is the same,

$$\langle \alpha jm|\mathbf{T}_1|\alpha' j' m'\rangle = \lambda(\alpha\alpha' jj')\langle \alpha jm|\mathbf{T}_2|\alpha' j' m'\rangle, \tag{71}$$

where α, α' are abbreviations for the rest of the quantum numbers needed to describe a state and λ is a constant for given $\alpha\alpha' jj'$, independent of m or m'.[15] Should λ be zero (or infinity) because of a selection rule on α or j for one of the operators, the theorem is of no use for those values of α and j. Since $\mathbf{j}$ is itself of type T we have, for matrix elements diagonal in α and j,

$$\langle \alpha jm|\mathbf{T}|\alpha jm'\rangle = \lambda(\alpha j)\langle \alpha jm|\mathbf{j}|\alpha jm'\rangle. \tag{72}$$

Returning now to our consideration of hyperfine interactions, let us compute the first-order perturbation due to $\mathscr{H}^{(1)}$, Eq. (49), on an atomic state $|nljm_j Im_I\rangle$. We consider first a non-s state ($l > 0$) so that the delta

---

[13] For a discussion of the relation between angular momenta and rotation operators see Feenberg and Pake (1953), p.12ff.

[14] These rules are taken from Griffith (1961), p.35.

[15] The proof of this theorem is given in Condon and Shortley (1935), Griffith (1961), and Feenberg and Pake (1953).

function term of (49) has an expectation value of zero and only the orbital and dipolar parts contribute. The operator (49) is in this case effectively

$$\mathscr{H}^{(1)} \to \mathscr{H}_L^{(1)} + \mathscr{H}_D^{(1)} = -gg_I\mu_0\mu_N\mathbf{I}\cdot\mathbf{T} \tag{73}$$

where

$$\mathbf{T} \equiv r^{-3}(\mathbf{l} - \mathbf{s}) + 3r^{-5}(\mathbf{s}\cdot\mathbf{r})\mathbf{r}. \tag{74}$$

It may be verified from the composition rules (1)–(3) that (74) is of type T with respect to **j**. Because of the complete separability of nuclear and electronic degrees of freedom in zeroth order we have, for a diagonal matrix element of $\mathscr{H}^{(1)}$,

$$\langle nljm_jIm_I|\mathscr{H}^{(1)}|nljm_jIm_I\rangle = -gg_I\mu_0\mu_N\langle Im_I|\mathbf{I}|Im_I\rangle\cdot\langle nljm_j|\mathbf{T}|nljm_j\rangle. \tag{75}$$

By the replacement theorem (72)

$$\langle nljm_j|\mathbf{T}|nljm_j'\rangle = \lambda(nlj)\langle nljm_j|\mathbf{j}|nljm_j'\rangle. \tag{76}$$

To evaluate the proportionality constant λ we take the scalar product of each side of (76) with $\langle nljm_j'|\mathbf{j}|nljm_j\rangle$ and sum over all m_j':

$$\begin{aligned}\sum_{m_j'=-j}^{j} &\langle nljm_j|\mathbf{T}|nljm_j'\rangle\cdot\langle nljm_j'|\mathbf{j}|nljm_j\rangle \\ &= \lambda(nlj)\sum_{m_j'=-j}^{j}\langle nljm_j|\mathbf{j}|nljm_j'\rangle\cdot\langle nljm_j'|\mathbf{j}|nljm_j\rangle.\end{aligned} \tag{77}$$

Since matrix elements of **j** off-diagonal in n, l or j vanish, the summations may be replaced by $\Sigma_{n'l'j'm_j'}$ over all the interior indices. By the rules for matrix multiplication we have then

$$\langle nljm_j|\mathbf{T}\cdot\mathbf{j}|nljm_j\rangle = \lambda(nlj)\langle nljm_j|j^2|nljm_j\rangle. \tag{78}$$

The matrix element on the right is $j(j+1)$, while

$$\mathbf{T}\cdot\mathbf{j} = \{r^{-3}(\mathbf{l} - \mathbf{s}) + 3r^{-5}(\mathbf{s}\cdot\mathbf{r})\mathbf{r}\}\cdot(\mathbf{l}\cdot\mathbf{s}) = r^{-3}(l^2 - s^2) + 3r^{-5}(\mathbf{s}\cdot\mathbf{r})^2 \tag{79}$$

since $\mathbf{r}\cdot\mathbf{l} = 0$. A diagonal matrix element of $\mathbf{T}\cdot\mathbf{j}$ is given by

$$\langle nljm_j|\mathbf{T}\cdot\mathbf{j}|nljm_j\rangle = \langle nlj|r^{-3}|nlj\rangle l(l+1) \tag{80}$$

since $3\langle|(\mathbf{s}\cdot\mathbf{r})^2/r^2|\rangle = \frac{3}{4}$, using (37), exactly cancelling $\langle|s^2|\rangle$.

From (78) and (80) we have

$$\lambda(nlj) = \langle nlj|r^{-3}|nlj\rangle\frac{l(l+1)}{j(j+1)} \tag{81}$$

so that, by (75) and (76),

$$\langle|\mathscr{H}^{(1)}|\rangle = a\langle|\mathbf{I}\cdot\mathbf{j}|\rangle \tag{82}$$

with

$$a(nlj) = -gg_I\mu_0\mu_N\langle nlj|r^{-3}|nlj\rangle \frac{l(l+1)}{j(j+1)}. \tag{83}$$

This formula was originally derived by Breit (1930, 1931a) and by Casimir (1930; see also Goudsmit, 1931).

The mathematical procedure leading to matrix elements of $\mathscr{H}^{(1)}$ may be given a simple physical interpretation in terms of the classical vector model for coupling of angular momenta. The magnetic field produced by the electrons at the nucleus is given, in terms of the operator (74), by $g\mu_0\mathbf{T}$. Since $\mathbf{j}$ is a conserved quantity for the atom, $\mathbf{l}$, $\mathbf{s}$, and $\mathbf{r}$, hence the magnetic field $g\mu_0\mathbf{T}$ may be considered to precess about $\mathbf{j}$. The effective average of one of these precessing vectors is equal to its component along $\mathbf{j}$ and we replace $\mathbf{T}$, for example, by $\langle|\mathbf{T}\cdot\mathbf{j}|\rangle\mathbf{j}/\langle|j^2|\rangle$. This simple substitution gives the same result, (82) and (83), as our more rigorous derivation.

$\mathbf{I}$ and $\mathbf{j}$ do not separately commute with the perturbed Hamiltonian so that m_I and m_j are no longer constants of the motion: however their sum

$$\mathbf{f} = \mathbf{I} + \mathbf{j} \tag{84}$$

does remain a conserved quantity. Evaluating $\langle|\mathbf{I}\cdot\mathbf{j}|\rangle$ in the f, m_f representation we find

$$E'(f) = \langle|\mathscr{H}^{(1)}|\rangle = \tfrac{1}{2}a[f(f+1) - j(j+1) - I(I+1)] \tag{85}$$

where $f = j + I,\ j + I - 1, \ldots, |j - I|$. The spacing between successive hyperfine levels is given by

$$E'(f) - E'(f-1) = af. \tag{86}$$

The energy level scheme (85) closely resembles the multiplet splitting of an atomic term, the spacings being in ratios of successive integers, and (86) is completely analogous to Landé's interval rule.

For a hydrogenlike atom (Bethe and Salpeter, 1957, p. 17)

$$\langle nl|r^{-3}|nl\rangle = \frac{Z^3}{a_0^3 n^3 l(l+\frac{1}{2})(l+1)}, \qquad l > 0. \tag{87}$$

Thus the interval factor a as an explicit function of n, l, and j is given by

$$a(nlj) = \frac{-gg_I\mu_0\mu_N Z^3}{a_0^3 n^3 (l+\frac{1}{2}) j(j+1)}, \qquad l > 0, \quad j = l \pm \tfrac{1}{2}. \tag{88}$$

Note that for given Z, a decreases rapidly with increasing n and l.

For a one-electron atom in an S state, the orbital and dipolar parts of $\mathcal{H}^{(1)}$ average to zero and only the delta function term contributes:

$$\mathcal{H}^{(1)} \to \mathcal{H}_c^{(1)} = -gg_I\mu_0\mu_N \frac{8\pi}{3} \delta(\mathbf{r})\mathbf{I}\cdot\mathbf{s}. \tag{89}$$

We arrive immediately for this case at an interval rule analogous to (85) and (86) but with $j = s = \frac{1}{2}$ and f limited to $I \pm \frac{1}{2}$. The interval factor for S states is given by

$$a = -gg_I\mu_0\mu_N \frac{8\pi}{3} |\psi_{n0}(0)|^2, \tag{90}$$

which was first derived by Fermi (1930). For solutions of the nonrelativistic Schrödinger equation we have

$$|\psi_{n0}(0)|^2 = \frac{Z^3}{\pi a_0^3 n^3} \tag{91}$$

and

$$a = -\frac{8}{3}\frac{gg_I\mu_0\mu_N Z^3}{a_0^3 n^3} \tag{92}$$

which is exactly what one would obtain by setting $l = 0$ and $j = \frac{1}{2}$ in (88) [even though (88) as derived above does not apply to this case]. Thus (88) gives, for all values of m, l, and j, the interval factor for one-electron atoms.[16] In the s electron case there are only two hyperfine levels, the doublet separation being given by

$$\Delta E = E'(I + \tfrac{1}{2}) - E'(I - \tfrac{1}{2}) = a\frac{2I + 1}{2}. \tag{93}$$

The hyperfine splitting in the $1S_{1/2}$ ground state of the hydrogen atom has received much consideration, both experimental and theoretical. From (92) and (93) we find, with $Z = 1$, $n = 1$, $I = \frac{1}{2}$, $g_I = 5.585486$, that

$$\Delta E = |a| = \tfrac{8}{3}gg_I\mu_0\mu_N a_0^{-3}. \tag{94}$$

ΔE is just the energy difference between the states of parallel and antiparallel relative orientations of the proton and electron spins. The transition frequency has been measured to extremely high accuracy by atomic

[16] The correctness of the $l > 0$ interval formula for s states is not fortuitous but is a consequence of the particular way in which the operator (28) was split up [cf. Eq. (38)ff]. A unified derivation of the Fermi and Breit-Casimir formulas may be carried out if the above decomposition is avoided and matrix elements of $\mathcal{H}^{(1)}$, Eq. (28), are evaluated directly.

beam (Prodell and Kusch, 1952, 1955; Wittke and Dicke, 1954, 1956) and optical polarization (L. W. Anderson *et al.*, 1960a) techniques. The best experimental value (Anderson *et al.*, 1960a) is

$$\Delta E/h = 1420.405726 \pm 0.000030 \text{ Mc/sec.}$$

If we compute $\Delta E/h$ using Eq. (94) with $g = -2$ we obtain 1421.11 Mc/sec. The most important correction we must make is for the finite mass of the nucleus (Breit and Meyerott, 1947, 1949) by replacing m by μ, the reduced mass, in a_0. This introduces a factor $(m/\mu)^{-3} = (1 + m/M)^{-3} = 0.9983679$ and gives $\Delta E/h = 1418.79$ Mc/sec. Additional small corrections have been made for (*i*) approximations to the Dirac theory (see Breit, 1930, 1931b)—i.e., neglect of the components of Ψ_2 and use of nonrelativistic wave function, (*ii*) radiative corrections (Kroll and Pollock, 1951, 1952; Karplus and Klein, 1952a,b)—virtual emission and absorption of photons, etc., and (*iii*) the effects of the proton's finite structure (Bechert and Meixner, 1935; Breit and Brown, 1948; Breit *et al.*, 1949; F. Low and Salpeter, 1950, 1951; Arnowitt, 1953; Newcomb and Salpeter, 1955; Moellering *et al.*, 1955). These latter effects are, however, much too small to explain the ~1.5 Mc/sec discrepancy, which is far greater than the experimental uncertainty. This discrepancy led Breit (1947) and Schwinger (1948, 1949) to suggest that the g factor of the electron's spin was not exactly -2, as predicted by the Dirac theory.[17] Calculation of the anomaly by quantum electrodynamics gives the following value for g, up to terms of the second order in α:[18,19,20]

$$g = -2\left(1 + \frac{\alpha}{2\pi} - \frac{0.328}{\pi^2}\alpha^2 + \cdots\right). \tag{95}$$

---

[17] The gyromagnetic factor for orbital motion (-1) is not affected by this correction. In applications of Eqs. (39) or (49), g multiplying the orbital term should be read as -2.

[18] The fourth-order (α^2) term was first computed by Karplus and Kroll (1950). An improved calculation is given by Sommerfield (1957).

[19] The anomalous g factor belongs to the category of radiative corrections, i.e., corrections to the Dirac theory by quantum electrodynamics. The origin of the anomaly may be described as follows. According to quantum electrodynamics, the electron undergoes virtual emission and reabsorption of photons. Photons of energy $\geq 2mc^2$ (~1 Mev) may, in turn, produce virtual positron-electron pairs. Thus there is always a small concentration of positrons and electrons around the "bare" electron. These share the spin angular momentum and hence contribute to the magnetic moment. Since the positively charged particles are attracted toward the core the contribution of the virtual electrons to the magnetic moment will predominate and hence the anomaly will tend to enhance the total magnetic moment.

[20] A recent high-precision measurement of the free-electron g factor gives, for the fourth-order term, $-(0.327 \pm 0.005)\alpha^2/\pi^2$ (Wilkinson and Crane, 1963).

Using this value of g in (94) along with the reduced mass correction we obtain

$$\Delta E/h = 1420.42 \quad \mathrm{Mc/sec},$$

in excellent agreement with experiment.

V. Hyperfine Structure of Complex Atoms and Ions

To derive expressions analogous to (39) and (49) for a many-electron system one should ideally begin with the many-particle analog of the Dirac equation and carry out a reduction to the nonrelativistic limit. Unfortunately, however, an exact relativistic wave equation, even for the two-electron case, cannot be written in closed form.[21] A major difficulty involves replacement of the $e^2 r_{ij}^{-1}$ Coulomb repulsions by operators which transform properly under Lorentz transformations. It is possible, however, at least in principle, to obtain approximations to various orders in the fine-structure constant (or in αZ for heavy atoms). It may be proven for the two-electron case,[22] and seems reasonable in general, that to first order in α the coupling between the nuclear magnetic moment and the electronic moments may be represented by a sum of operators like (39) or (49), one for each electron. We may thus generalize, for a complex atom, that

$$\mathcal{H}^{(1)} = \sum_{k=1}^{N} \mathcal{H}_k^{(1)} = -g g_I \mu_0 \mu_N \mathbf{I}\cdot \sum_{k=1}^{N} \Big\{ [r_k^{-3}(\mathbf{l}_k - \mathbf{s}_k) + 3r_k^{-5}(\mathbf{s}_k\cdot\mathbf{r}_k)\mathbf{r}_k]_{r_k>0} + \frac{8\pi}{3}\,\delta(\mathbf{r}_k)\mathbf{s}_k \Big\} \tag{96}$$

where the summation index k runs over the labels of the N electrons. We shall assume henceforth that (96) is the correct many-electron hyperfine interaction operator. It can easily be verified that the entire summation is an operator of type T with respect to the total electronic angular momentum $\mathbf{J}$ where

$$\mathbf{J} = \sum_{k=1}^{N} \mathbf{j}_k = \sum_{k=1}^{N} (\mathbf{l}_k + \mathbf{s}_k). \tag{97}$$

---

[21] Generalizations of Dirac's relativistic equation are considered by Breit (1929, 1932a,b, 1938); Eddington (1929); Gaunt (1929a,b); Oppenheimer (1930); Salpeter and Bethe (1951); Brown (1952). An excellent discussion is given in Bethe and Salpeter (1957), Sect. 38ff.

[22] This is implicit in the reduction carried out in Bethe and Salpeter (1957), Sect. 39.

Thus we have, in complete analogy to (85) and (86), the hyperfine structure interval rule

$$E'(F) = \langle|\mathscr{H}^{(1)}|\rangle = A\langle|\mathbf{I}\cdot\mathbf{J}|\rangle = \tfrac{1}{2}A[F(F+1) - J(J+1) - I(I+1)]$$

and

$$E'(F) - E'(F-1) = AF \tag{98}$$

where

$$\mathbf{F} = \mathbf{J} + \mathbf{I} \tag{99}$$

and

$$F = J + I, J + I - 1, \ldots, |J - I|.$$

The $^6S_{5/2}$ ground state of Mn atom ($I = 5/2$) provides a good illustration of this interval structure (White and Ritschl, 1930) (see Fig. 2). The average energy of the six hyperfine levels, weighted by their $(2F+1)$-fold M_F degeneracies, is equal to the unperturbed energy. This follows from the invariance of the trace of an operator under change of representation, for

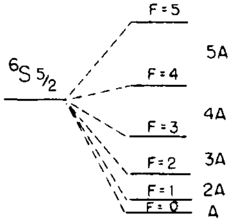


Fig. 2. Hyperfine of the $3d^54s^2\ ^6S_{5/2}$ ground term of Mn^{55} ($I = 5/2$) from data of White and Ritschl, (1930). The interval factor A is about 0.002 cm^{-1}.

$$\begin{aligned}\sum_{F=|J-I|}^{J+I}\ \sum_{M_F=-F}^{F} E'(F) &= A\sum_{F}\sum_{M_F}\langle FM_F|\mathbf{I}\cdot\mathbf{J}|FM_F\rangle = A\ \mathrm{tr}\ \mathbf{I}\cdot\mathbf{J}\\ &= A\sum_{M_J=-J}^{J}\ \sum_{M_I=-I}^{I}\langle M_JM_I|\mathbf{I}\cdot\mathbf{J}|M_JM_I\rangle\\ &= A\sum_{M_J}\ \sum_{MI} M_IM_J = 0.\end{aligned} \tag{100}$$

There is a very close analogy between the Landé interval rule for atomic multiplet splitting in the Russell–Saunders scheme and the hfs interval formula (98)—the quantum numbers L, S, and J being replaced by J, I, and F, respectively. This is exceedingly useful, since formulas for fine-structure energy spacings, selection rules, and relative intensities can be taken over without modification. For example, the selection rule on F for optical transitions in the absence of an external magnetic field is $\Delta F = 0, \pm 1$, excepting $F = 0 \rightarrow F = 0$, exactly the same as the intermultiplet selection rule on J. The hfs of an optical transition may exhibit directly the level scheme of Eq. (98)—characterized by spacings in ratios of successive integers—provided that the splitting of the initial or of the final state is disproportionately small, or zero (as in the case $J = 0$). The hfs of various lines of the praseodymium spectrum (White, 1929) (Fig. 3) provides an excellent illustration of interval structure. These so-called flag patterns are made up of several components decreasing in intensity as well as interval toward either shorter or longer wavelengths. The relative intensities of the hyperfine components may be deduced theoretically (Hill, 1929). They reflect principally the statistical weights $(2F + 1)$ of the levels.

Let us presume, for the moment, that the flag pattern is due to the hyperfine structure of the higher-energy electronic state. When the lowest-frequency component corresponds to the smallest F value (e.g., lines in the upper half of Fig. 3), the interval constant A is positive and the hyperfine structure is termed "normal." In the opposite case, A is negative and the hfs is "inverted." Evidently, both types of structure can occur in the same spectrum.

The origin of normal and inverted hfs may be understood from a classical viewpoint. It is related to the signs of the terms in (49) or (96). Orbital motion of an electron produces a magnetic field at the nucleus antiparallel to $\mathbf{l}$. For a nucleus having $g_I > 0$, the most stable orientation of the nuclear moment is that parallel to the field, hence antiparallel to $\mathbf{l}$. This corresponds to a positive coupling constant multiplying $\mathbf{I}\cdot\mathbf{l}$ and thus to a normal contribution to hfs. In the dipolar interaction, the most stable orientation of $\mathbf{S}$ with respect to $\mathbf{I}$ is the parallel one. Thus the dipolar part contributes to inverted hfs. The contact part is again normal since it has the same sign as the orbital term. This holds rue, however, only if the electron spin polarization at the nucleus has the same sense as the total spin of the atom. It may be surmised that, in the case of interepnetrating magnetic moments, the fundamental circulatory nature of the moments manifests itself, so that their coupling resembles somewhat the orbital interaction. The foregoing discussion applies *mutatis mutandis* when the flag pattern is due to the lower state or when g_I is negative.

As an illustration, we consider the hfs of a ^2P atomic term. In the $^2P_{1/2}$ level $\mathbf{L}$ and $\mathbf{S}$ are antiparallel; thus the orbital and dipolar contributions reinforce one another, one being normal the other inverted in any

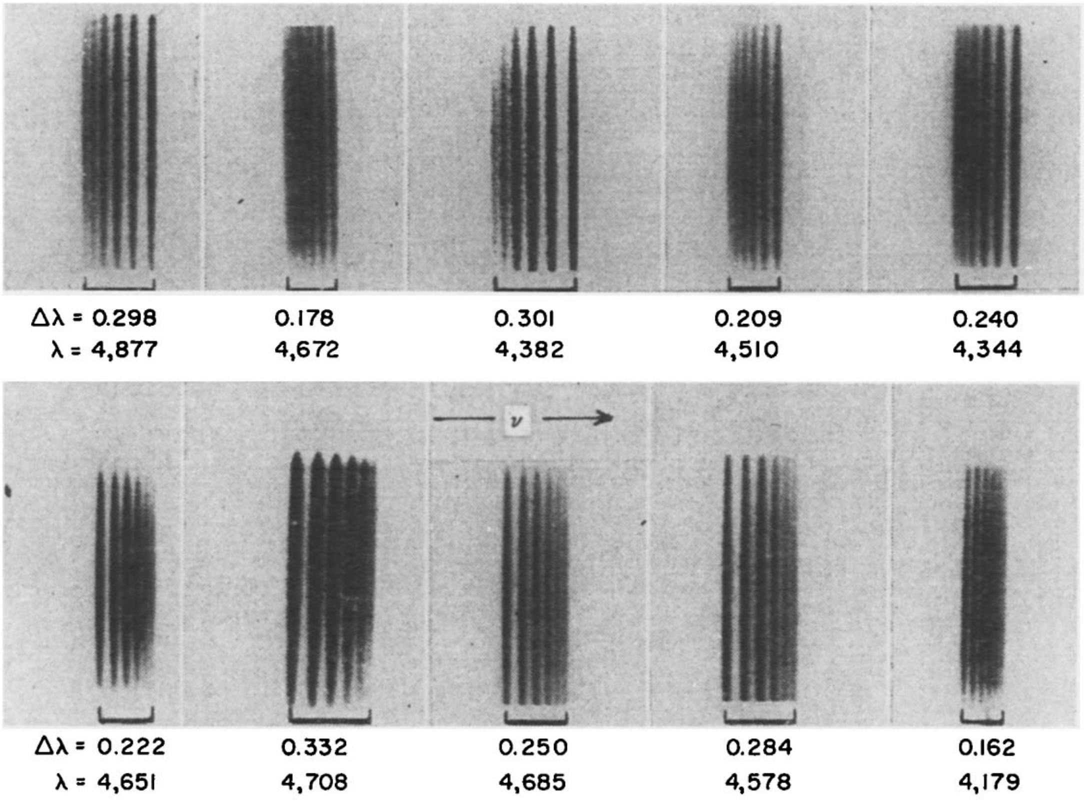


Fig. 3. Hyperfine structure of several lines of the praseodymium spectrum (after White, 1929). The interval structures probably all come from atomic levels of large J, so that the multiplicity is $2I+1$ with $I = 5/2$.

case. The corresponding ${}^2P_{3/2}$ level, however, shows partial cancellation, so that

$$|A({}^2P_{1/2})| > |A({}^2P_{3/2})|.$$

provided only that the influence of the contact term is small.

It should not be presumed that the neat interval patterns illustrated above occur in the hfs of all spectral lines. The foregoing must be modified under the following circumstances:

(*i*) When initial and final states have comparable splittings
(*ii*) When there is appreciable coupling with the nuclear quadrupole moment
(*iii*) When the element is a mixture of isotopes
(*iv*) When quantization of J breaks down
(*v*) When an external magnetic field is applied.

The influence of a magnetic field (Zeeman effect) is treated extensively in the literature (see, for example, Ramsey, 1955, Chapter 3). We shall consider the remaining effects in the next few paragraphs.

Nuclei with spin $I > \frac{1}{2}$ may have an electric quadrupole moment and couple to the gradient of the electric field produced by the atomic electrons. The resulting energy levels may be written (see, for example, Casimir, 1936)

$$E'_Q(F, J, I) = \frac{B[\frac{3}{4}K(K+1) - I(I+1)J(J+1)]}{2I(2I-1)J(2J-1)} \tag{101}$$

where $K = F(F+1) - J(J+1) - I(I+1)$ and B is known as the quadrupole coupling constant. B depends on both the nuclear quadrupole moment and the electric field gradient. Deviation of the hyperfine levels from interval-rule spacing occurs because of the $K(K+1)$ factor in (101). For light nuclei we generally have $A \gg B$ so that any deviations are slight. From the middle of the periodic table onward, it is not uncommon to find B sizably larger than A, so that the hfs of many heavy elements is determined principally by quadrupole coupling. In a few cases, the influence of a nuclear magnetic octupole moment has also been observed (Jaccarino *et al.*, 1954; Kusch and Eck, 1954).

When there are several isotopes present the observed hyperfine structure will be complicated by their overlapping contributions (see Fig. 4, for example). In addition to simple superposition of lines arising from the different values of I and g_I among the isotopes, there may be small shifts in the optical transition frequencies. These "isotope shifts" (Hughes and Eckart, 1930; Pauli and Peierls, 1931; Bartlett, 1931; Bartlett and Gibbons,

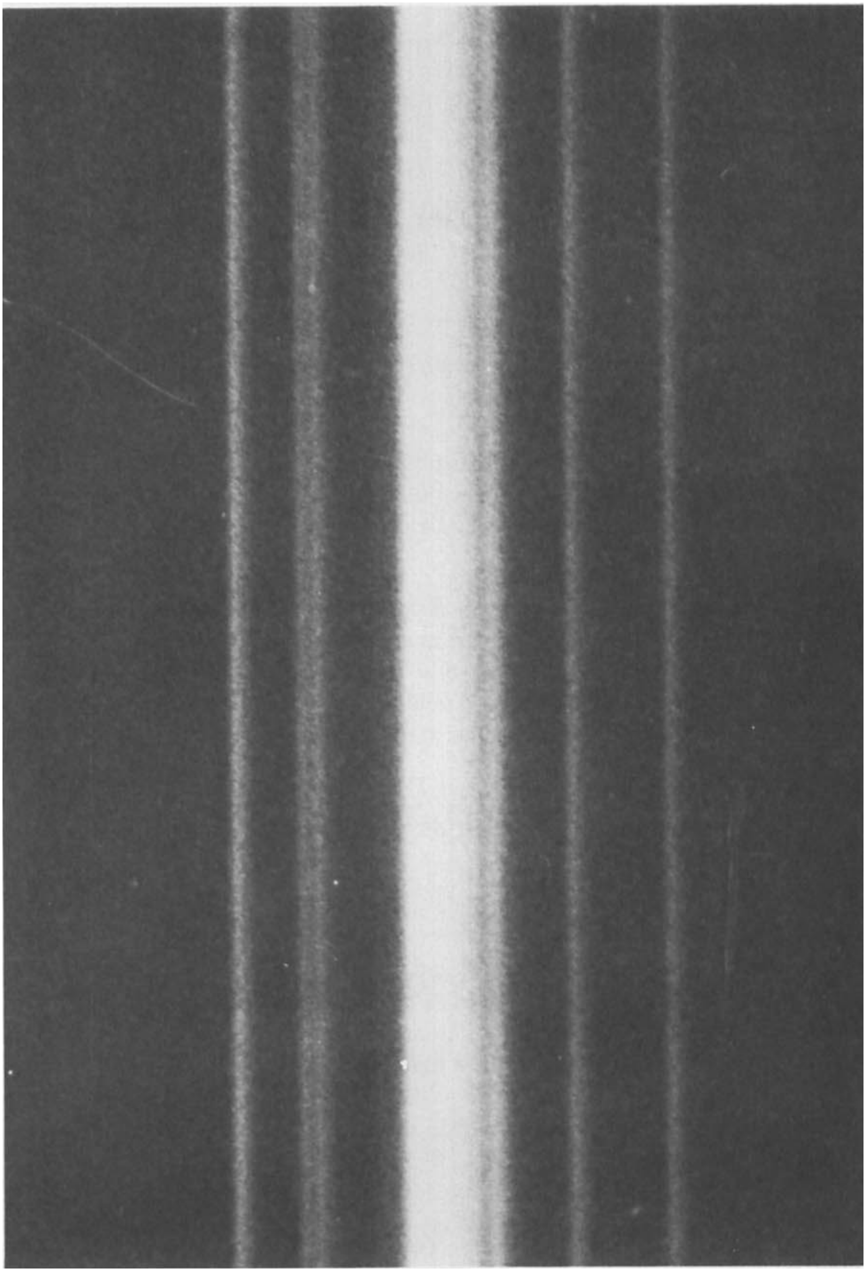


Fig. 4. Hyperfine structure of the 4358 *A* line of mercury. The light central region is due to the even (spinless) isotopes. The hyperfine splitting arises from the isotopes of mass number 199 ($I=\frac{1}{2}$) and 201 ($I=\frac{3}{2}$).

1933)[23] are due in light atoms to the effect of noninfinite nuclear mass—the reduced mass correction plus a coupling term $-(\hbar^2/M)\Sigma_{i<j}\nabla_i\cdot\nabla_j$, which gives a contribution varying with electronic state. In heavy atoms, variations in nuclear size cause additional isotope shifts. Isotopic mixing is, of course, not a fundamental difficulty and can be eliminated by working with isotopically pure samples or, better yet, by employing atomic beam techniques.

Derivation of the interval rule (98) rests on the supposition that J is a good quantum number, i.e., that hyperfine structure is much smaller than fine structure. If this is the case, only matrix elements diagonal in J need be considered. If the two types of interaction are of comparable magnitude, however, J is no longer conserved. Only F remains a constant of the motion and we must consider mixing of nearby states of different J. Such an effect has been observed in the $1s2p^3P$ term of He^3 atom (Fred *et al.*, 1951). Due to some fortuitous cancellation, the spin-orbit coupling is very weak (Bethe and Salpeter, 1957, Sect. 40) and is, in fact, exceeded by the

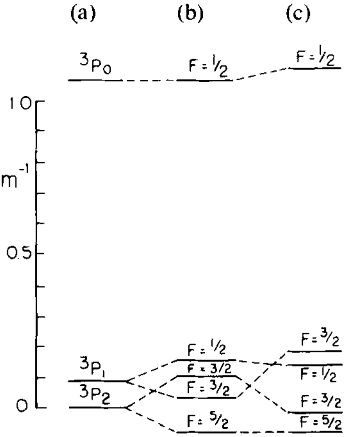


Fig. 5. Energy level scheme in the $1s2p^3P$ term of He^3 (after Fred *et al.*, 1951). (*a*) Fine structure (observed He^4); (*b*) hyperfine structure calculated from matrix elements diagonal in J; (*c*) hyperfine structure allowing for perturbations (observed He^3). The hfs is inverted because of the negative nuclear g factor. The fine structure is also inverted.

hyperfine coupling. The energy level scheme is shown in Fig. 5. The unperturbed (by hyperfine interaction) $^3P_2$ and $^3P_1$ levels are separated by less than 0.1 cm^{-1} and are about 1 cm^{-1} from the $^3P_0$ level, as can be

---

[23] A review of isotope effects is given in Kopfermann (1958), Sect. 33–37.

inferred from the He^4 spectrum. The $F = 3/2$ states arising from $J = 2$ and $J = 1$ ($I = \frac{1}{2}$ for He^3) interact very strongly and are shifted apart by several tenths of a wave number. The $F = \frac{1}{2}$ states from $J = 1$ and $J = 0$ are more weakly perturbed, whereas the $F = 5/2$ state is not perturbed at all, there being no other nearby level of the same F. Similar interactions have been observed in the $1s3p\,{}^3P$ and $1s3d\,{}^3D$ terms of He^3 as well as in the isoelectronic Li^7(II) ion (Güttinger and Pauli, 1931).

It should be abundantly clear from the preceding discussion that assignment of hyperfine coupling constants from optical spectra is often a complex matter, quite aside from difficulties in resolution. Atomic beam and paramagnetic resonance techniques are inherently superior in this respect since all transitions occur within the manifold of a single electronic level. Still, the main advantage of magnetic resonance methods is their resolving power. Whereas in optical spectroscopy hyperfine coupling is a second-order effect and resolution of the order of 0.001 cm^{-1} is about the best one can do, in magnetic resonance sensitivity of $10^{-6}\ \text{cm}^{-1}$ ($\sim$0.01 Mc/sec) is readily attainable. The selection rules for magnetic dipole transitions among hyperfine levels in zero magnetic field are the same as those for electric dipole transitions, $\Delta F = 0, \pm 1$, except $0 \rightarrow 0$. More important, however, in magnetic resonance work are transitions in a magnetic field, which we shall not discuss here.

Before going on to the theoretical evaluation of hyperfine interval constants in many-electron atoms, let us consider briefly the two-electron case, for which rather good results can be obtained by a simple computation. He^3 in its $1s^2\,{}^1S$ ground state and in its $1s2s\,{}^1S$ excited state shows no hfs, of course, since $J = 0$. The $1s2s\,{}^3S$ state, however, exhibits a rather large splitting of 6739.71 ± 0.05 Mc/sec (Weinreich and Hughes, 1954). Let us take the following approximation to the wave function of the ${}^3S$ state

$$\Psi(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{\sqrt{2}}[\psi_{1s}(r_1)\psi_{2s}(r_2) - \psi_{2s}(r_1)\psi_{1s}(r_2)]\begin{cases} \alpha(1)\alpha(2) \\ \frac{1}{\sqrt{2}}[\alpha(1)\beta(2) + \beta(1)\alpha(2)] \\ \beta(1)\beta(2) \end{cases} \tag{102}$$

The contact interaction operator for two electrons is given by

$$\mathscr{H}_c^{(1)} = -gg_I\mu_0\mu_N\frac{8\pi}{3}\{\delta(\mathbf{r}_1)\mathbf{I}\cdot\mathbf{s}_1 + \delta(\mathbf{r}_2)\mathbf{I}\cdot\mathbf{s}_2\}. \tag{103}$$

Assuming that the orbitals ψ_{1s} and ψ_{2s} are normalized and mutually orthogonal, we find for the expectation value of the delta function

$$\langle|\delta(\mathbf{r}_1)|\rangle = \langle|\delta(\mathbf{r}_2)|\rangle = \tfrac{1}{2}[|\psi_{1s}(0)|^2 + |\psi_{2s}(0)|^2]. \tag{104}$$

Hence with $\mathbf{s}_1 + \mathbf{s}_2 = \mathbf{S}$ we have

$$\langle|\mathscr{H}_C^{(1)}|\rangle = A\langle|\mathbf{I}\cdot\mathbf{S}|\rangle \tag{105}$$

in terms of the interval constant

$$A = -gg_I\mu_0\mu_N \frac{8\pi}{3} \tfrac{1}{2}[|\psi_{1s}(0)|^2 + |\psi_{2s}(0)|^2]. \tag{106}$$

If we further approximate ψ_{1s} and ψ_{2s} by hydrogenlike functions with $Z = 2$ and $Z = 1$, respectively, and neglect their slight nonorthogonality, we obtain, using (91),

$$|\psi_{1s}(0)|^2 = \frac{8}{\pi a_0^3}, \qquad |\psi_{2s}(0)|^2 = \frac{1}{8\pi a_0^3}. \tag{107}$$

We see that the $1s$ contribution to the Fermi term far outweighs the $2s$. The large density at the nucleus of the unpaired $1s$ orbital accounts for the large observed hyperfine splitting. Substituting numerical values in (106) and (107) with $g(\mathrm{He}^3) = -4.255088$ we find

$$A \approx -4400 \quad \mathrm{Mc/sec}.$$

The He^3 hfs is inverted because of the negative nuclear g-factor. By the interval rule (98)

$$\Delta E = |E'(F = \tfrac{3}{2}) - E'(F = \tfrac{1}{2})| = \tfrac{3}{2}A \tag{108}$$

and $\Delta E/h \approx 6600$ Mc/sec, in remarkable agreement with experiment. Using a twelve-term Hylleraas function and making corrections for mass polarization and anomalous electron g factor, the theoretical value

$$\frac{\Delta E}{h} = 6739.84 \quad \mathrm{Mc/sec}$$

has been obtained (Traub and Foley, 1958).[24] Analogous computations have been carried out for the $1s2s\,^3S$ state of $\mathrm{Li}^7(\mathrm{II})$ (Luke *et al.*, 1952).

General formulas for the interval factor A in many-electron atoms showing Russell–Saunders (L-S) coupling were first worked out by means of sum rules by Goudsmit (1931). These results have been rederived

---

[24] See also Teutsch and Hughes (1954). Relativistic and radiative corrections are considered by Sessler and Foley (1955).

and extended by Trees (1953; see also Hubbs *et al.*, 1958) and by Schwartz (1955) using Racah's tensor formalism (Racah, 1942, 1943; Fano and Racah, 1959). The influence of intermediate coupling (between *L-S* and *j-j*) on hyperfine structure formulas has also been examined (Goudsmit, 1931; Breit and Wills, 1933; Wybourne, 1962). Breakdown of the Russell–Saunders coupling scheme becomes a possibility whenever there are two or more electrons outside of closed shells. Deviations are especially important in the transition metal and the rare earth series and become extreme (about halfway between *L-S* and *j-j*) in the actinide elements. We will not consider the most general formulation for complex atoms, which is extremely complicated. Instead, we will content ourselves with a few special cases which are more amenable to explicit treatment.[25]

In the Russell–Saunders coupling approximation, the observables $\mathscr{H}$, L^2, S^2, J^2, and J_z form a complete compatible set for a complex atom. J^2 and J_z are preferable to the equivalent pair L_z, S_z since the former apply even after spin-orbit coupling is introduced. Atomic states (exclusive of the nuclear spin variables) are then represented by the sets $|ELSJM\rangle$. By the replacement theorem (72) we have the following relation for A:

$$A(ELSJ)\langle ELSJM|\mathbf{J}|ELSJM'\rangle = -gg_I\mu_0\mu_N\left\langle ELSJM\left|\sum_{k=1}^{N}\left\{[r_k^{-3}(\mathbf{l}_k - \mathbf{s}_k) + 3r_k^{-5}(\mathbf{s}_k\cdot\mathbf{r}_k)\mathbf{r}_k]_{r_k>0} + \frac{8\pi}{3}\,\delta(\mathbf{r}_k)\mathbf{s}_k\right\}\right|ELSJM'\right\rangle. \quad (109)$$

Since A is independent of M and M', we can simplify the computation by taking $M = M' = J$. We need only consider the z component of (109). Thus

$$AJ = -gg_I\mu_0\mu_N\left\langle ELSJJ\left|\sum_{k=1}^{N}\left\{[r_k^{-3}(l_{kz} - s_{kz}) + 3r_k^{-5}(\mathbf{s}_k\cdot\mathbf{r}_k)z_k]_{r_k>0} + \frac{8\pi}{3}\,\delta(\mathbf{r}_k)s_{kz}\right\}\right|ELSJJ\right\rangle. \quad (110)$$

The operator in (110) is most readily evaluated in the M_LM_S representation. The appropriate transformation leads to

$$AJ = \sum_{M_LM_SM_L'M_S'}\langle LSJJ|LSM_LM_S\rangle\left\langle ELSM_LM_S\left|\sum_{k=1}^{N}\cdots\right|ELSM_L'M_S'\right\rangle \times \langle LSM_L'M_S'|LSJJ\rangle \quad (111)$$

---

[25] Formulas for hyperfine coupling constants in some special cases are considered by Breit and Doermann (1930); Breit (1931b); Goudsmit (1933); Breit and Wills (1933); Fermi and Segre (1933a,b); Casimir (1936).

where the $\langle LSJM|LSM_LM_S\rangle$ are Wigner coefficients. In order to evaluate the central matrix elements it is necessary to have an approximation to the atomic wave function. The easiest to handle is a Hartree–Fock determinantal function. Let us consider first the level of highest multiplicity, $J = L + S$, within a Russell–Saunders term, i.e., ${}^4S_{3/2}$, ${}^2P_{3/2}$, ${}^2D_{5/2}$, but not ${}^2P_{1/2}$, ${}^2D_{3/2}$, etc. The only nonvanishing Wigner coefficient is then

$$\langle L, S, L + S, L + S|LSLS\rangle = 1, \tag{112}$$

so that (111) reduces to

$$AJ = \left\langle ELSLS \,\middle|\, \sum_{k=1}^{N} \cdots \,\middle|\, ELSLS \right\rangle. \tag{113}$$

For a single-determinant wave function, the sum over electrons reduces to a sum over occupied orbitals. If the usual central-field approximation is made for the one-electron functions we have then

$$AJ = -gg_I\mu_0\mu_N \sum_{\substack{\text{occupied}\\ nlm_lm_s}} \Big\{\langle nl|r^{-3}|nl\rangle[m_l + \langle lm_l|3\cos^2\theta - 1|lm_l\rangle m_s] + \frac{8\pi}{3}\langle nl|\delta(\mathbf{r})|nl\rangle\Big\}. \tag{114}$$

By various theorems on spherical harmonics it can be shown that

$$\langle lm_l|3\cos^2\theta - 1|lm_l\rangle = \frac{2l(l + 1) - 6m_l^2}{(2l + 3)(2l - 1)}. \tag{115}$$

Noting that sums over complete l subshells cancel, i.e.,

$$\sum_{m_l=-l}^{l} m_l = 0, \qquad \sum_{m_l=-l}^{l} \langle lm_l|3\cos^2\theta - 1|lm_l\rangle = 0, \tag{116}$$

it is easy to see that, for a closed-shell atom ($J = L = S = 0$), the interval factor is zero, quite aside from the vanishing of hyperfine splitting. Atoms and ions in multiplet S states such as:

N: $1s^22s^22p^3\ {}^4S_{3/2}$
P: $[\mathrm{Ne}]3s^23p^3\ {}^4S_{3/2}$
Mn^{++}: $[\mathrm{Ar}]3d^5\ {}^6S_{5/2}$
Mn: $[\mathrm{Ar}]3d^54s^2\ {}^6S_{5/2}$
Eu: $[\mathrm{Xe}]4f^7\ {}^8S_{7/2}$
Am: $[\mathrm{Rn}]5f^77s^2\ {}^8S_{7/2}$

would also be expected to show no hyperfine structure. The open shells are spherically symmetrical by Unsöld's theorem and have parallel spins by Hund's rules. For the above S states the summations (116) again cancel out the orbital and dipolar terms. The contact part should also vanish since the unpaired spins reside in orbitals ($l > 0$) with zero density at the nucleus. Contrary to expectation, however, multiplet S atoms do show sizable hyperfine splittings.[26]

One possible explanation of the above anomaly might be spin-orbit coupling which, in the case of nitrogen, for example, would admix some ${}^2P_{3/2}$ character into the ${}^4S_{3/2}$ ground state. Interval factors computed on the basis of this mechanism are, however, too small by about two orders of magnitude. A more reasonable explanation is configuration interaction (Goudsmit, 1931; Fermi and Segre, 1933a,b; Goudsmit and Bacher, 1933; Abragam, 1950; Abragam and Pryce, 1951a,b).

One method of improving upon a single determinant wave function is by making a linear combination of determinants. Each determinant may be thought to represent a configuration, and the improvement in the description of an atomic state is ascribed to "configuration interaction" (CI). Hyperfine structure in multiplet S atoms comes about because of admixture of configurations in which an s electron is promoted to a higher orbital. For example, in nitrogen atom we would have

$$1s^22s^22p^3 \leftrightarrow 1s^22s2p^33s, \quad \text{etc.}$$

The unpairing of an s subshell results in a net spin density at the nucleus and a large contribution to the contact term provided that the two unpaired s electrons recouple to form a triplet. The excited configuration must, however, remain of ${}^4S_{3/2}$ symmetry over-all in order to mix with the ground configuration. A more precise description of the excited configuration is:

$$1s^22s3s\ {}^3S_1,\ 2p^3\ {}^4S_{3/2},\ {}^4S_{3/2},$$

which has as its representation a sum of determinantal functions with suitably permuted spins. Configuration interaction may also augment the hfs of atoms already having open-shell structures, e.g.,

$$\text{Li: } 1s^22s \leftrightarrow 1s2s^2,\ 1s2s3s, \quad \text{etc.}$$

$$\text{O: } 1s^22s^22p^2 \leftrightarrow 1s^22s2p^23s, \quad \text{etc.}$$

In the case of lithium, about 30% of the hyperfine splitting is contributed by CI.

---

[26] A survey of hfs data for S state ions is given in W. Low (1960), Tables XXVI–XXVIII.

Koster (1952) invoked configuration interaction to explain certain irregularities in the hfs of gallium. An explicit computation of the "*s* electron effect" in the first transition series was attempted by Abragam *et al.* (1955). They considered only one excited configuration and found a coupling constant too small by an order of magnitude. It is now understood that a large number of configurations, including the continuum, make comparable contributions, so that admixture of a single excited configuration is not sufficient. More successful application of the CI procedure has been made possible by the advent of electronic computers. Nesbet's (1960) CI calculation on lithium atom predicts a hyperfine splitting 98.8% the experimental value. CI calculations by Bessis and colleagues on nitrogen, oxygen, and boron have also been in fairly good agreement with experiment (see Bessis-Mazloum and Lefebvre-Brion, 1960; Bessis *et al.*, 1961, 1962, 1963; also Bessis, 1962).

It should be noted that, although intermediate coupling and configuration interaction may be very important in determining the magnitude of the interval factor A, they do not affect the validity of the interval rule (98)—in contrast to the F-F mixing we considered earlier (e.g., ^3P He^3).

The most successful treatment of the *s* electron effect to date has been through the approach of spin polarization (also known as exchange polarization or core polarization). It was pointed out by Slater (1951a) and others (Pople and Nesbet, 1954; Berthier, 1954; Nesbet, 1955; Löwdin, 1955b) that the conventional Hartree–Fock formalism under the equivalence restriction (same radial function for each n, l group, independent of m_l or m_s) neglects important differences in exchange interactions. This is immediately evident from the form of the integrodifferential equations for orbital functions in the Hartree–Fock scheme (see, for example, Hartree, 1957):

$$\begin{aligned}
&\left\{-\frac{\hbar^2}{2m}\nabla^2 - \frac{Ze^2}{r} + \sum_{l\neq k} e^2 \int d\tau'\, \phi_l^*(\mathbf{r}')|\mathbf{r}-\mathbf{r}'|^{-1}\phi_l(\mathbf{r}')\right\}\phi_k(\mathbf{r}) \\
&\qquad = \varepsilon_k\phi_k(\mathbf{r}) + \sum_{\substack{l\neq k\\ \text{same spin}}} e^2\left\{\int d\tau'\phi_l^*(\mathbf{r}')|\mathbf{r}-\mathbf{r}'|^{-1}\phi_k(\mathbf{r}')\right\}\phi_l(\mathbf{r}).
\end{aligned} \tag{117}$$

The last summation contains only exchange integrals with orbitals l having the same spin as k. In lithium atom, for example, if the $2s$ orbital is associated with spin α, there will be exchange interaction with the $1s\alpha$ orbital but none with the $1s\beta$. Under the equivalence restriction, the $1s$ orbital function is constrained to satisfy an equation like (117) with the *average* of the $1s\alpha$ and $1s\beta$ exchange integrals (i.e., half of the $1s\alpha$)

substituted in place of the last sum. Only by introducing configuration interaction, which will mix unequally configurations such as $1s\alpha 2s\alpha ns\beta$ and $1s\beta 2s\alpha ns\alpha$, can exchange interactions be accounted for. However, a multi-determinant CI function is more difficult to handle computationally. In order to retain the convenience of a single determinant and still allow for different exchange interactions it was suggested (Pople and Nesbet, 1954; Berthier, 1954; Nesbet, 1955) that the equivalence restriction on m_s be dropped. The formalism allowing different orbitals for different m_s values within the Hartree–Fock scheme is most often known as the "unrestricted Hartree–Fock" (UHF) method (Nesbet, 1955). The terminology is somewhat loose in that the m_l restriction and r, θ, ϕ separability of the orbital functions are still maintained. A more precise nomenclature would be "spin-polarized Hartree–Fock" (SPHF).[27]

By introducing additional freedom into the functional form of the orbitals—$1s\alpha$ and $1s\beta$ in lithium, for example, being no longer constrained to have the same radial dependence—we cannot fail in accordance with the variational principle to improve upon our wave function. The effect leading to different orbitals for different spins is called exchange polarization since it is caused by differences in exchange interactions. The imbalance thereby induced between α and β orbitals of each s shell will exhibit itself as a nonzero spin density (hence the alternative designations spin polarization and core polarization). In particular, the spin densities at the nucleus of the $ns\alpha$ and $ns\beta$ electrons will no longer exactly cancel and there will be a net contribution to the contact term given by

$$\langle ns\alpha|\delta(\mathbf{r})s_z|ns\alpha\rangle + \langle ns\beta|\delta(\mathbf{r})s_z|ns\beta\rangle = \tfrac{1}{2}[|\psi_{ns\alpha}(0)|^2 - |\psi_{ns\beta}(0)|^2]. \quad (118)$$

The importance of spin polarization in the calculation of magnetic properties of atoms and solids has been discussed by Slater (1951a), Sternheimer (1952), and others (Wood, 1957; Cohen and Reif, 1957).[28]

The first attempts to calculate polarization contributions to hfs were made using empirical exchange potentials (Kohn, 1954; Jones and Schiff, 1954; Kjeldaas and Kohn, 1956; Wood and Pratt, 1957; Heine, 1957) suggested by Seitz (1935) and by Slater (1951b). These calculations were successful only in the qualitative sense—in being able to predict the direction and approximate magnitude of the spin polarization effects.

---

[27] For other discussions of modified Hartree–Fock schemes see Lefebvre (1953); Chirgwin (1957); McWeeny (1960); Löwdin (1960); Nesbet (1961).

[28] Computations of magnetic properties (other than hyperfine splittings) using spin-polarized wave functions have been made by Freeman and Watson (1960a,b) and Watson and Freeman (1960, 1961a).

Exchange interactions among electrons—apart from those already accounted for by antisymmetry—involve considerably less energy than the classical electrostatic repulsions. It was suggested by Cohen *et al.* (1959) that the exchange terms be treated as a perturbation on the restricted self-consistent field orbitals. Their results for hyperfine splittings in lithium and sodium atoms were quite encouraging. The relatively minute extent to which the s orbitals are polarized lends credence to the perturbation approach. An analogous procedure has been applied to calculation of the hfs of atomic nitrogen (Blinder, 1960b; Das and Mukherji, 1960). Goodings (1961), however, has criticized the perturbation approach to spin polarization, pointing out that, despite the small magnitude of the corrections, a small set of perturbation basis functions cannot adequately represent the subtle behavior of the atomic wave function near the nucleus.

The most direct approach to spin polarization, obviating the inaccuracies inherent in exchange potentials or in a perturbation expansion, is, of course, an *ab initio* application of the self-consistent field procedure with the equivalence restriction removed at the outset. Unrestricted Hartree–Fock calculations of the hyperfine splitting in lithium by Sachs (1960) and Goodings (1961) predict about 97% of the experimental value. In contrast, the restricted HF solution gives only the $2s$ contribution, which amounts to about 70% of the observed splitting. Applications of the UHF method of heavier atoms by Goodings (1961) and by Freeman and Watson (1961, 1962; Watson and Freeman, 1961b,c) have yielded fair to good results for hfs. Not surprisingly, accuracy tends to decrease with increasing atomic number. Results for nitrogen (Bessis *et al.*, 1961; Blinder, 1960b; Das and Mukherji, 1960; Goodings, 1961; Freeman and Watson, 1961) have been erratic because of almost exact cancellation of the $1s$ and $2s$ contributions to spin polarization at the nucleus.

A spin-polarized determinantal function, e.g., $|1s\alpha 1s'\beta 2s\alpha|$ for lithium, is not an exact eigenfunction of S^2 because of the unpairing of the s shell orbitals.[29] Sachs (1960) found, for example, that his lithium UHF function had $\langle|S^2|\rangle = 0.75001573$, a minute deviation from pure doublet character (for which $\langle|S^2|\rangle = 3/4$). In other UHF calculations, larger deviations have been found, the largest thus far being in the case of nitrogen (Freeman and Watson, 1961), $\langle|S^2|\rangle = 3.7579$ (vs 3.75 for pure quartet). This presents a serious conceptual difficulty in the UHF method since the nonrelativistic, spin-independent Hamiltonian commutes with S^2 and therefore all eigenfunctions of $\mathscr{H}$ should be simultaneously eigenfunctions of S^2. Even

---

[29] A detailed treatment of the spin degeneracy problem in the UHF method is given by Pratt (1956).

a variational approximation to an eigenfunction should be a pure spin state since off-diagonal matrix elements in S^2 vanish.

There is, however, a procedure to restore the correct spectroscopic symmetry. Formally the UHF function may be represented by a linear combination of pure spin states. For lithium, introducing the sum and difference of the $1s$ and $1s'$ functions

$$\begin{aligned} u &= 1s + 1s' \\ v &= 1s - 1s' \end{aligned} \tag{119}$$

we find that $|1s\alpha 1s'\beta 2s\alpha|$ breaks up into four determinants. Ignoring multiplicative constants, these are

$$\Phi_1 = |u\alpha u\beta 2s\alpha|, \tag{120a}$$

$$\Phi_2 = |v\alpha v\beta 2s\alpha|, \tag{120b}$$

and the symmetrical combination

$$2^{-1/2}\{|u\alpha v\beta 2s\alpha| + |u\beta v\alpha 2s\alpha|\}. \tag{120c}$$

The functions Φ_1, Φ_2 represent pure doublet spin states, analogous to restricted HF determinants. The first of these represents essentially the ground state which would be obtained under the equivalence restriction; the second is a doubly excited state of the same symmetry. The sum (120c) may be written

$$(\tfrac{1}{3})^{1/2}\Phi_3 + (\tfrac{2}{3})^{1/2}\Phi_4 \tag{121}$$

with

$$\Phi_3 = 6^{-1/2}\{|u\alpha v\beta 2s\alpha| + |u\beta v\alpha 2s\alpha| - 2|u\alpha v\alpha 2s\beta|\} \tag{121a}$$

and

$$\Phi_4 = 3^{-1/2}\{|u\alpha v\beta 2s\alpha| + |u\beta v\alpha 2s\alpha| + |u\alpha v\alpha 2s\beta|\}. \tag{121b}$$

Φ_3 is still another doublet function, for a singly excited state; Φ_4 is, however, a component of a quartet spin state and it is this part which accounts for the deviation of the UHF function from pure spin character. The UHF solution is thus of the form

$$\Psi_{\mathrm{UHF}} = |1s\alpha 1s'\beta 2s\alpha| = a_1\Phi_1 + a_2\Phi_2 + a_3\Phi_3 + a_4\Phi_4 \tag{122}$$

where Φ_1, Φ_2, and Φ_3 are doublets, Φ_4 is a quartet, and the a's are numerical coefficients depending on overlap integrals among the functions $1s$, $1s'$, and $2s$.

The quartet part of (122) may be removed by applying a spin projection operator (Löwdin, 1955b) to the determinantal function Ψ_{UHF}. An alternative method is given by Heine (1957). The general form of a spin projection operator is

$$\Lambda(S_i) = \prod_{j \neq i} \left\{ \frac{S^2 - S_j(S_j + 1)}{S_i(S_i + 1) - S_j(S_j + 1)} \right\}. \tag{123}$$

The effect of (123) on a wave function is to remove all but the part corresponding to spin quantum number S_i—or, in the usual terminology, to "project out" the S_i component. For lithium the doublet projection operator reduces to a single factor

$$\Lambda_{\text{doublet}} = \frac{S^2 - \frac{15}{4}}{\frac{3}{4} - \frac{15}{4}} \tag{124}$$

and has the effect

$$\Lambda_{\text{doublet}} \Psi_{\text{UHF}} = a_1\Phi_1 + a_2\Phi_2 + a_3\Phi_3. \tag{125}$$

Making use of the identity

$$S^2 = S_z^2 + S_z + S_- S_+ \tag{126}$$

where

$$S_\pm = S_x \pm i\, S_y = \sum_{k=1}^{3} s_{k\pm}, \tag{127}$$

it follows that

$$\Lambda_{\text{doublet}} |1s\alpha 1s'\beta 2s\alpha| = -\tfrac{1}{3}\{|1s\alpha 1s'\alpha 2s\beta| + |1s\beta 1s'\alpha 2s\alpha| - 2|1s\alpha 1s'\beta 2s\alpha|\}, \tag{128}$$

which is the explicit form of (125). The projected function (128) is easily seen to be a spin doublet in which the 1*s* and 2*s* spins are internally coupled in a triplet. The effect of the projection operator on the single determinant function (122) has been to mix in two other determinants differing by spin flips. Although the core polarization is clearly reduced by projection, we note that it is not entirely eliminated—an important conclusion, for otherwise the UHF method would be entirely invalid (see Pratt, 1956). To illustrate, Sachs (1960) found that projection decreased the calculated hyperfine splitting for lithium from 97.2 to 88.3% of the experimental value. Freeman and Watson (1961) for nitrogen found a reduction from about three times to about twice the correct value.

In the procedure leading to the function (125) or (128), spin projection is carried out *after* the optimization of the Hartree–Fock orbitals. Whereas this projected UHF function has the advantage of being a spin eigenfunction, it suffers from the drawback of not having been determined directly from the variational principle. The unprojected UHF function (122) has just the reverse attributes. A superior approach, suggested by Löwdin

(1955a), would be to project *first* and then carry out the optimization of the orbitals. This is known as the extended Hartree–Fock method (EHF) and is, at present, beyond computational capabilities. Conceptually, at least, the EHF function for lithium would be of the form

$$\Psi_{\mathrm{EHF}} = \zeta_1\Phi_1 + \zeta_2\Phi_2 + \zeta_3\Phi_3 \tag{129}$$

analogous to (125) except that the constants $a_1, \ldots$ would be replaced by variational parameters $\zeta_1, \ldots$ and the orbital functions $1s$, $1s'$, $2s$ in $\Phi_1, \ldots$ would be determined so as to optimize the entire projected function (129) rather than the single determinant (122).

The apparent success of the unprojected UHF method for hyperfine splittings presents somewhat of a paradox. First, why does removal of the quartet component usually give a worse answer even though the projected wave function is fundamentally more correct? Second and more puzzling, allowing different orbitals for different spins picks up only a minute fraction of the correlation energy[30]—1% at best—while far more important contributions to electron correlation are ignored. Why then is the spin density improved so markedly over the restricted Hartree–Fock result?

The first question has been discussed by Marshall (1961) and by Heine (1963). They show that, if certain exchange integrals are small relative to configurational promotion energies, the spin density given by Ψ_{EHF}, presumably the best over-all wave function, is equal in first order to the spin density given by Ψ_{UHF} rather than that for the projected function $\Lambda\Psi_{\mathrm{UHF}}$. Lefebvre (1959)[31] reached essentially the same conclusion and, in addition, identified Ψ_{UHF} with the CI function involving all singly excited determinants (also not an eigenfunction of S^2). It would seem then that the more imperative criterion on a wave function is that it be determined in its final form by the variational principle—this taking

---

[30] We cite some illustrative values of the total energy in lithium atom calculations (experimental value $E = -7.4783$ a.u.). Nesbet (1960), with limited configuration interaction, finds a decrease in the total energy from $E_{\mathrm{HF}} = -7.431765$ to $E_{\mathrm{CI}} = -7.431849$. The unrestricted Hartree–Fock calculation of Sachs (1960) shows an even more minute improvement: $E_{\mathrm{HF}} = -7.432727$ a.u. vs $E_{\mathrm{UHF}} = -7.432751$ a.u. In an independent series of calculations on lithium, Nesbet and Watson (1960) report the following total energy values:

$E_{\mathrm{HF}} = -7.432723$ a.u.,
$E_{\mathrm{UHF}} = -7.432747$ a.u.,
$E_{\mathrm{ProjUHF}} = -7.432764$ a.u.

[31] See also Bessis–Mazloum and Lefebvre–Brion (1960); Bessis *et al.* (1961, 1962, 1963); Bessis (1962).

precedence over spectroscopic purity. The result for lithium (Sachs, 1960) supports the conclusion that Ψ_{UHF} is better than $\Lambda\Psi_{\text{UHF}}$ whereas that for nitrogen (Freeman and Watson, 1961) is obviously inconclusive.

Heine (1963) has also cast some light on the second question above—on the relation of exchange polarization to correlation in general. He has shown that the function Φ_3 [Eq. (121a)] represents pure exchange polarization, leading to unbalanced $1s$ spin density but contributing nothing in first order to the correlation energy. Φ_2 [Eq. (120b)], on the other hand, represents in-out correlation in the $1s^2$ shell. It contributes a major part of the correlation energy but nothing in first order to the spin density. The perturbations represented by Φ_2 and Φ_3 do not "interfere" since they are drawn from different excited "configurations"—$vv2s$ and $uv2s$, respectively. Thereby we may rationalize the fact that correct spin densities are given by functions which are no better energetically than the Hartree–Fock.

A more fundamental method of introducing correlation is by explicit inclusion of r_{ij} terms into the wave function (see, for example, Löwdin, 1959; Slater, 1960, Chapter 18) along lines initiated by Hylleraas and by James and Coolidge for two-electron system. A recent calculation on lithium by Berggren and Wood (1963) using strongly correlated wave functions, has yielded superior results for both the energy (−7.47630 a.u.) and the hyperfine splitting (98.8 % of experimental). No asymmetry or polarization of the core appears anywhere in this calculation. This will perhaps be the ultimate method of treating hyperfine interactions in complex systems.

ACKNOWLEDGMENT

This paper is based on a series of lectures given at Harvard University during the Fall Semester, 1962. I am deeply grateful to Professor W. A. Klemperer for his support and encouragement during this time.

REFERENCES

ABRAGAM, A. (1950). *Phys. Rev.* **79**, 534.

ABRAGAM, A., and PRYCE, M. H. L. (1951a). *Proc. Roy. Soc.* **A205**, 135.

ABRAGAM, A., and PRYCE, M. H. L. (1951b). *Proc. Roy. Soc.* **A206**, 164.

ABRAGAM, A., HOROWITZ, J., and PRYCE, M. H. L. (1955). *Proc. Roy. Soc.* **A230**, 169.

ANDERSON, L. W., PIPKIN, F. M., and BAIRD, J. C. (1958). *Phys. Rev. Letters* **1**, 229.

ANDERSON, L. W., PIPKIN, F. M., and BAIRD, J. C. (1959). *Phys. Rev.* **116**, 87.

ANDERSON, L. W., PIPKIN, F. M., and BAIRD, J. C. (1960a). *Phys. Rev. Letters* **4**, 69.

ANDERSON, L. W., PIPKIN, F. M., and BAIRD, J. C. (1960b). *Phys. Rev.* **120**, 1279.

ARNOWITT, R. (1953). *Phys. Rev.* **92**, 1002.

BACK, E., and GOUDSMIT, S. (1927). *Z. Physik* **43**, 321.

BACK, E., and GOUDSMIT, S. (1928). *Z. Physik* **47**, 174.

BARTLETT, J. H. (1931). *Nature* **128**, 408.

Bartlett, J. H., and Gibbons, J. J. (1933). *Phys. Rev.* **44**, 538.

Bechert, K., and Meixner, J. (1935). *Ann. Physik* [5] **22**, 525.

Berggren, K. F., and Wood, R. F. (1963). *Phys. Rev.* **130**, 198.

Beringer, R., and Heald, M. A. (1954). *Phys. Rev.* **95**, 1474.

Berthier, G. B. (1954). *J. Chim. Phys.* **51**, 363.

Bessis, N. (1962). Doctoral Thesis, University of Paris.

Bessis, N., Lefebvre-Brion, H., and Moser, C. M. (1961). *Phys. Rev.* **124**, 1124.

Bessis, N., Lefebvre-Brion, H., and Moser, C. M. (1962). *Phys. Rev.* **128**, 213.

Bessis, N., Lefebvre-Brion, H., and Moser, C. M. (1963). *Phys. Rev.* **130**, 1441.

Bessis-Mazloum, N., and Lefebvre-Brion, H. (1960). *Compt. Rend.* **251**, 648.

Bethe, H. A. (1933). *In* "Handbuch der Physik" (H. Geiger and K. Scheel, eds.), 2nd ed., Vol. 24, Part I. Springer, Berlin.

Bethe, H. A., and Salpeter, E. E. (1957). "Quantum Mechanics of One- and Two-Electron Atoms." Academic Press, New York.

Blinder, S. M. (1960a). *J. Mol. Spectry.* **5**, 17.

Blinder, S. M. (1960b). *Bull. Am. Phys. Soc.* [2] **5**, 14.

Bohr, A., and Weisskopf, V. F. (1950). *Phys. Rev.* **77**, 94.

Bowers, K. D., Kamper, R. A., and Knight, R. B. D. (1957). *J. Sci. Instr.* **34**, 49.

Breit, G. (1929). *Phys. Rev.* **34**, 553.

Breit, G. (1930). *Phys. Rev.* **35**, 1447.

Breit, G. (1931a). *Phys. Rev.* **37**, 51.

Breit, G. (1931b). *Phys. Rev.* **38**, 463.

Breit, G. (1932a). *Phys. Rev.* **36**, 383.

Breit, G. (1932b). *Phys. Rev.* **39**, 616.

Breit, G. (1938). *Phys. Rev.* **53**, 153.

Breit, G. (1947). *Phys. Rev.* **72**, 984.

Breit, G., and Brown, G. E. (1948). *Phys. Rev.* **74**, 1278.

Breit, G., and Doermann, F. W. (1930). *Phys. Rev.* **36**, 1732.

Breit, G., and Meyerott, R. (1947). *Phys. Rev.* **72**, 1023.

Breit, G., and Meyerott, R. (1949). *Phys. Rev.* **75**, 1447.

Breit, G., and Wills, L. A. (1933). *Phys. Rev.* **44**, 470.

Breit, G., Brown, G. E., and Arfken, G. B. (1949). *Phys. Rev.* **76**, 1299.

Brix, P. (1952). *Z. Physik.* **132**, 579.

Brossel, J., and Kastler, A. (1949). *Compt. Rend.* **229**, 1213.

Brossel, J., Sagalyn, P. H., and Bitter, F. (1950). *Phys. Rev.* **79**, 225.

Brown, G. E. (1952). *Phil. Mag.* [7] **43**, 467.

Casimir, H. B. G. (1930). Cited in Pauling and Goudsmit (1930), p.208ff.

Casimir, H. B. G. (1936). "On the Interaction Between Atomic Nuclei and Electrons." Teyler's Tweede Genootschap, Haarlem.

Chirgwin, B. H. (1957). *Phys. Rev.* **107**, 1013.

Cohen, M. H., and Reif, F. (1957). *Solid State Phys.* **5**, 322.

Cohen, M. H., Goodings, D. A., and Heine, V. (1959). *Proc. Phys. Soc.* (*London*) **73**, 811.

Condon, E. U., and Shortley, G. J. (1935). "Theory of Atomic Spectra." Cambridge Univ. Press, London and New York.

Crawford, M. F., and Schawlow, A. L. (1949). *Phys. Rev.* **67**, 1310.

Das, T. P., and Mukherji, A. (1960). *J. Chem. Phys.* **33**, 1808.

Dehmelt, H. G. (1955). *Phys. Rev.* **99**, 527.

DIRAC, P. A. M. (1928). *Proc. Roy. Soc.* **A117**, 610.

DIRAC, P. A. M. (1958). "Principles of Quantum Mechanics," 4th ed. Oxford Univ. Press, London and New York.

EDDINGTON, A. S. (1929). *Proc. Roy. Soc.* **A122**, 358.

FABRY, C., and PEROT, A. (1897). *Ann. Chim. Phys.* (*Paris*) **12**, 459.

FANO, U., and RACAH, G. (1959). "Irreducible Tensorial Sets." Academic Press, New York.

FEENBERG, E., and PAKE, G. E. (1953). "Notes on the Quantum Theory of Angular Momentum." Addison-Wesley, Reading, Massachusetts.

FERMI, E. (1930). *Z. Physik* **60**, 320.

FERMI, E., and SEGRE, E. (1933a). *Z. Physik* **82**, 729.

FERMI, E., and SEGRE, E. (1933b). *Mem. Reale Accad. Ital.* **4**, 131.

FERRELL, R. A. (1960). *Am. J. Phys.* **28**, 484.

FOCK, V. (1926a). *Z. Physik* **38**, 242.

FOCK, V. (1926b). *Z. Physik.* **39**, 226.

FONER, S. N., JEN, C. K., COCHRAN, E. L., and BOWERS, V. A. (1958). *J. Chem. Phys.* **28**, 351.

FONER, S. N., COCHRAN, E. L., BOWERS, V. A., and JEN, C. K. (1960). *J. Chem. Phys.* **32**, 963.

FRED, M., TOMPKINS, F. S., BRODY, J. K., and HAMERMESH, M. (1951). *Phys. Rev.* **82**, 406.

FREEMAN, A. J., and WATSON, R. E. (1960a). *Phys. Rev.* **118**, 1168.

FREEMAN, A. J., and WATSON, R. E. (1960b). *Phys. Rev. Letters* **5**, 498.

FREEMAN, A. J., and WATSON, R. E. (1961). *MIT Solid State Mol. Theory Group Quart. Progr. Rept.* No. 40.

FREEMAN, A. J., and WATSON, R. E. (1962). *Phys. Rev.* **127**, 2058.

FROSCH, R. A., and FOLEY, H. M. (1952). *Phys. Rev.* **88**, 1337.

GAUNT, J. A. (1929a). *Phil. Trans. Roy. Soc.* **228**, 151.

GAUNT, J. A. (1929b). *Proc. Roy. Soc.* **A122**, 513.

GOODINGS, D. A. (1961). *Phys. Rev.* **123**, 1706.

GORDON, W. (1926). *Z. Physik* **40**, 117.

GOUDSMIT, S. (1931). *Phys. Rev.* **37**, 663.

GOUDSMIT, S. (1933). *Phys. Rev.* **43**, 636.

GOUDSMIT, S., and BACHER, R. F. (1929). *Phys. Rev.* **34**, 1501.

GOUDSMIT, S., and BACHER, R. F. (1933). *Phys. Rev.* **43**, 894.

GRIFFITH, J. S. (1961). "The Theory of Transition Metal Ions." Cambridge Univ. Press, London and New York.

GÜTTINGER, P., and PAULI, W. (1931). *Z. Physik* **67**, 743.

HARGREAVES, J. (1929). *Proc. Roy. Soc.* **A124**, 568.

HARGREAVES, J. (1930). *Proc. Roy. Soc.* **A127**, 407.

HARTREE, D. R. (1957). "The Calculation of Atomic Structures." Wiley, New York.

HEALD, M. A., and BERINGER, R. (1954). *Phys. Rev.* **96**, 654.

HEINE, V. (1957). *Phys. Rev.* **107**, 1002.

HEINE, V. (1963). *Czech. J. Phys.* **B13**, 619.

HILL, E. L. (1929). *Proc. Natl. Acad. Sci. U.S.* **15**, 779.

HOLLOWAY, W. W., and NOVICK, R. (1958). *Phys. Rev. Letters* **1**, 507.

HUBBS, H. C., MARRUS, R., NIERENBERG, W. A., and WORCESTER, J. L. (1958). *Phys. Rev.* **109**, 390.

HUGHES, D. S., and ECKART, C. (1930). *Phys. Rev.* **36**, 694.

JACCARINO, V., KING, J. G., SATTEN, R. A., and STROKE, H. H. (1954). *Phys. Rev.* **94**, 1798.

JEN, C. K., FONER, S. N., COCHRAN, E. L., and BOWERS, V. A. (1956). *Phys. Rev.* **104**, 846.

JEN, C. K., FONER, S. N., COCHRAN, E. L., and BOWERS, V. A. (1958). *Phys. Rev.* **112**, 1169.

JONES, H., and SCHIFF, B. (1954). *Proc. Phys. Soc.* (*London*) **A67**, 217.

KARPLUS, R., and KLEIN, A. (1952a). *Phys. Rev.* **85**, 972.

KARPLUS, R., and KLEIN, A. (1952b). *Phys. Rev.* **87**, 848.

KARPLUS, R., and KROLL, N. M. (1950). *Phys. Rev.* **77**, 536.

KJELDAAS, T., and KOHN, W. (1956). *Phys. Rev.* **101**, 66.

KLEIN, O. (1927). *Z. Physik* **41**, 407.

KOHN, W. (1954). *Phys. Rev.* **96**, 590.

KOPFERMANN, H. (1958). "Nuclear Moments" (Translation by E. E. Schneider). Academic Press, New York.

KOSTER, G. F. (1952). *Phys. Rev.* **86**, 148.

KROLL, N. M., and POLLOCK, F. (1951). *Phys. Rev.* **84**, 594.

KROLL, N. M., and POLLOCK, F. (1952). *Phys. Rev.* **86**, 876.

KUSCH, P., and ECK, T. G. (1954). *Phys. Rev.* **94**, 1799.

LAMBERT, R. H., and PIPKIN, F. M. (1962). *Phys. Rev.* **128**, 198.

LEFEBVRE, R. (1953). *Compt. Rend.* **237**, 1158.

LEFEBVRE, R. (1959). *Cahiers Phys.* **110**, 369 (1959).

LÖWDIN, P.-O. (1955a). *Phys. Rev.* **97**, 1474.

LÖWDIN, P.-O. (1955b). *Phys. Rev.* **97**, 1509.

LÖWDIN, P.-O. (1959). *Advan. Chem. Phys.* **2**, 207.

LÖWDIN, P.-O. (1960). *Rev. Mod. Phys.* **32**, 328.

LOW, F., and SALPETER, E. E. (1950). *Phys. Rev.* **77**, 361.

LOW, F., and SALPETER, E. E. (1951). *Phys. Rev.* **83**, 478.

LOW, W. (1960). *Solid State Phys.* Suppl. 2.

LUKE, P. J., MEYEROTT, R. E., and CLENDENIN, W. W. (1952). *Phys. Rev.* **85**, 401.

MCWEENY, R. (1960). *Rev. Mod. Phys.* **32**, 335.

MANDL, F. (1957). "Quantum Mechanics," 2nd ed. Academic Press, New York.

MARSHALL, W. (1961). *Proc. Phys. Soc.* (*London*) **A78**, 113.

MEGGERS, W. F., and BURNS, K. (1927). *J. Opt. Soc. Am.* **14**, 449.

MICHELSON, A. A. (1891). *Phil. Mag.* [5] **31**, 338.

MILFORD, F. J. (1960). *Am. J. Phys.* **28**, 521.

MOELLERING, W. M., ZEMACH, A. C., KLEIN, A., and LOW, F. E. (1955). *Phys. Rev.* **100**, 441.

NESBET, R. K. (1955). *Proc. Roy. Soc.* **A230**, 312.

NESBET, R. K. (1960). *Phys. Rev.* **118**, 681.

NESBET, R. K. (1961). *Rev. Mod. Phys.* **33**, 28.

NESBET, R. K., and WATSON, R. E. (1960). *Ann. Phys.* (*N.Y.*) **9**, 260.

NEWCOMB, W., and SALPETER, E. E. (1955). *Phys. Rev.* **97**, 1146.

OPPENHEIMER, J. R. (1930). *Phys. Rev.* **35**, 461.

PAULI, W. (1924). *Naturwiss.* **12**, 741.

PAULI, W., and PEIERLS, R. E. (1931). *Physik. Z.* **32**, 670.

PAULING, L., and GOUDSMIT, S. (1930). "Structure of Line Spectra." McGraw-Hill, New York.

PIPKIN, F. M., and LAMBERT, R. H. (1962). *Phys. Rev.* **127**, 787.

POPLE, J. A., and NESBET, R. K. (1954). *J. Chem. Phys.* **22**, 571.

PRATT, G. W. (1956). *Phys. Rev.* **102**, 1303.

PRODELL, A., and KUSCH, P. (1952). *Phys. Rev.* **88**, 184.

PRODELL, A., and KUSCH, P. (1955). *Phys. Rev.* **100**, 1188.

RACAH, G. (1931). *Z. Physik.* **71**, 431.

RACAH, G. (1932). *Nature*, **129**, 723.

RACAH, G. (1942). *Phys. Rev.* **62**, 438.

RACAH, G. (1943). *Phys. Rev.* **63**, 367.

RADO, G. T. (1962). *Am. J. Phys.* **30**, 716.

RAMSEY, N. F. (1955). "Molecular Beams." Oxford Univ. Press, London and New York.

ROSENTHAL, J., and BREIT, G. (1932). *Phys. Rev.* **41**, 459.

SACHS, L. M. (1960). *Phys. Rev.* **117**, 1504.

SALPETER, E. E., and BETHE, H. A. (1951). *Phys. Rev.* **84**, 1232.

SCHIFF, L. I. (1955). "Quantum Mechanics," 2nd ed. McGraw-Hill, New York.

SCHRÖDINGER, E. (1926). *Ann. Physik.* [4] **81**, 109.

SCHWARTZ, C. (1955). *Phys. Rev.* **97**, 380.

SCHWINGER, J. (1948). *Phys. Rev.* **73**, 416.

SCHWINGER, J. (1949). *Phys. Rev.* **76**, 790.

Seitz, F. (1935). *Phys. Rev.* **47**, 400.

SESSLER, A. M., and FOLEY, H. M. (1955). *Phys. Rev.* **98**, 6.

SESSLER, A. M., and MILLS, R. L. (1958). *Phys. Rev.* **110**, 1453.

SLATER, J. C. (1951a). *Phys. Rev.* **82**, 538.

SLATER, J. C. (1951b). *Phys. Rev.* **81**, 385.

SLATER, J. C. (1960). "Quantum Theory of Atomic Structure," Vol. 2. McGraw-Hill, New York.

SOMMERFIELD, C. M. (1957). *Phys. Rev.* **107**, 328.

STERNHEIMER, R. (1952). *Phys. Rev.* **86**, 313.

TEUTSCH, W. B., and HUGHES, V. W. (1954). *Phys. Rev.* **95**, 1461.

TRAUB, J., and FOLEY, H. M. (1958). *Phys. Rev.* **111**, 1098.

TREES, R. E. (1953). *Phys. Rev.* **92**, 308.

VAN DER WAERDEN, B. L. (1932). "Die Gruppentheoretische Methode in der Quantenmechanik." Springer, Berlin.

WATSON, R. E., and FREEMAN, A. J. (1960). *Phys. Rev.* **120**, 1125 and 1134.

WATSON, R. E., and FREEMAN, A. J. (1961a). *Phys. Rev. Letters* **6**, 277 and 388E.

WATSON, R. E., and FREEMAN, A. J. (1961b). *Phys. Rev.* **123**, 2027.

WATSON, R. E., and FREEMAN, A. J. (1961c). *J. Appl. Phys.* **32**, 118S.

WEINREICH, G., and HUGHES, V. W. (1954). *Phys. Rev.* **95**, 1451.

WHITE, H. E. (1929). *Phys. Rev.* **34**, 1397.

WHITE, H. E. (1934). "Introduction to Atomic Spectra." McGraw-Hill, New York.

WHITE, H. E., and RITSCHL, R. (1930). *Phys. Rev.* **35**, 1146.

WILKINSON, D. T., and CRANE, H. R. (1963). *Phys. Rev.* **130**, 852.

WITTKE, J. P., and DICKE, R. H. (1954). *Phys. Rev.* **96**, 530.

WITTKE, J. P., and DICKE, R. H. (1956). *Phys. Rev.* **103**, 620.

WOOD, J. H. (1957). *MIT Solid State Mol. Theory Group Quart. Progr. Rept.* No. 24.

WOOD, J. H., and PRATT, G. W. (1957). *Phys. Rev.* **107**, 995.

WYBOURNE, B. G. (1962). *J. Chem. Phys.* **37**, 1807.

The Theory of Pair-Correlated Wave Functions

R. McWEENY and E. STEINER*

Quantum Theory Group
University of Keele, Keele
Staffordshire, England

I. Introduction

Orbital approximations to the electronic wave functions of many-electron systems usually start from a single antisymmetrized product,[1]

$$\Psi_0 = \mathsf{A}[\psi_1(1)\psi_2(2) \cdots \psi_N(N)], \tag{I.1}$$

in which $\psi_1, \psi_2, \ldots, \psi_N$ are orthonormal spin-orbitals and are conveniently determined variationally, using self-consistent field (SCF) methods, to give a best one-configuration wave function. The shortcomings of the approximation are well-known, and arise mainly from its inability to describe any kind of correlation between electrons of unlike spin. To overcome this failure most directly the form (I.1) should be abandoned, as in the functions of Hylleraas (1929), James and Coolidge (1933), and others; but in general this leads to severe difficulties.

Since in practice it is often possible to distinguish different groups of electrons (e.g., in the shell model of the atom), correlation being large within a group but small between different groups, there is considerable

---

* Present address: Department of Chemistry, University of Exeter, England.

[1] When there is no risk of confusion we use i to indicate space-spin variables $x_i = r_i, s_i$ of particle i.

scope for an intermediate approximation of the form (McWeeny, 1959, 1960)

$$\Psi_0 = \mathsf{A}[\phi_A(1, \ldots, N_A)\phi_B(N_A + 1, \ldots, N_A + N_B) \cdots] \qquad (\text{I}.2)$$

in which each electron group has its own wave function, which need not itself be of product form and may therefore recognize correlation effects within the group. The "separated" form is an expression of the belief that intergroup correlation effects are small. As in SCF theory, best group functions $\phi_A, \phi_B, \ldots$ can be determined by variational methods (McWeeny, 1959, 1960), and there is evidence that when different groups are well-localized the form (I.2) may be highly accurate.

The one-configuration forms (I.1) and (I.2) may both be improved by the method of configuration interaction. "Excited" functions, based on a new configuration of orbitals or group functions, are allowed to mix with Ψ_0 and the mixing coefficients are determined variationally (or by perturbation theory) to give a many-configuration function which makes some allowance for correlation between the electrons of different orbitals or groups. With functions of type (I.2) this configuration interaction amounts to a recognition of "polarization" and "dispersion" effects between the groups (McWeeny, 1959, 1960), provided that the groups are spatially localized; otherwise the interpretation is somewhat more formal.

The separation of electrons into groups may or may not have a clear physical justification and, whereas configuration interaction based on orbital products of type (I.1) can in principle lead to an exact solution, that based on (I.2), with a fixed choice of groups, cannot. For these reasons Sinanoğlu (1961, 1962, 1964) has proposed a variant of the approach based on (I.1) in which correlation is introduced not by using an extended set of excited orbital configurations but by introducing *correlated pair functions* to describe electrons excited two at a time. Szasz (1959, 1960, 1962a,b, 1963) has used a closely related approach, deriving from first principles the matrix elements involving Ψ_0 and the excited configurations. Much of this analysis, however, is already implicit in the group function method and many of the results of Sinanoğlu and Szasz can be derived easily from the matrix element expressions for generalized products of the type (I.2). On the other hand, most of the work so far has been on closed-shell systems and the spin coupling problems which arise in other cases (when the correlated wave function is required to be an exact spin eigenfunction) do not appear to have been fully discussed.

The aim of this paper is to give a simple and unified treatment of correlated pair theories, paying particular attention to spin-coupling requirements, and to discuss various ways in which development might be made.

II. The Pair-Sea Formulation

We shall use a two-group wave function, in which one group describes a pair of electrons while the other describes a "sea" of (N-2) electrons in which the pair moves. The initial approximation is taken to be the antisymmetrical spin-orbital product of Hartree–Fock theory:

$$\Psi_0(1, 2, \ldots, N) = M_0 \mathsf{A}[\psi_A(1)\bar{\psi}_A(2) \cdots \psi_R(i) \cdots \psi_X(N)] \tag{II.1}$$

where the bar denotes a spin orbital with β spin factor. Any pair of spin-orbitals then defines an "unexcited pair," and refinement of the wave function will be achieved by admixture of functions in which each pair in turn has been excited. Different choices of group, in particular that corresponding to *single* excitations, will be considered in a later section.

The spin-orbitals in (II.1) will be written

$$\psi_R(i) = R(i)\alpha(i), \qquad \bar{\psi}_R(i) = R(i)\beta(i).$$

Letters $A, B, \ldots$ will denote the doubly occupied orbitals in Ψ_0; $U, V, \ldots$ will denote the singly occupied (all associated with α spin); and (R, S) will be used for a general pair. Ψ_0 is a spin eigenfunction with quantum numbers S, M $(M = S)$ when there are $2S$ orbitals in the group $U, V, \ldots$; and the "excited" functions to be used in the configuration interaction calculation should be chosen to have identical spin eigenvalues. With a spin-free Hamiltonian, it is of course sufficient to consider the case $M = S$, since the energy is independent of M.

We now consider individual pairs of electrons (e.g., those in spin-orbitals ψ_R, ψ_S) in order to introduce a correlation factor for each pair. To do this (II.1) may be written in an alternative form as

$$\Psi_0 = M' \mathsf{A}[\phi_0(1, 2)\Phi_0(3, \ldots, N)] \tag{II.2}$$

where ϕ_0 is a function for the pair selected and Φ_0 describes the "sea" of (N-2) electrons in which the pair moves. The electrons associated with ϕ_0 may always be denoted by 1, 2 since the product is fully antisymmetrized by the operator A; and M' may be chosen so that (II.1) and (II.2) are identical. However, for a given pair of spatial orbitals $R(i)$, $S(j)$ there will be in general various possible choices of ϕ_0 and Φ_0, depending on the coupling of the spins. If μ is used to denote a particular selection of orbitals, $\mu = (RS)$, the different couplings may be indicated by subscripts (s, m); thus, $\phi_0(\mu_{sm}|1, 2)$ will be used to denote a particular pair of spatial orbitals μ, chosen from the basic set in Ψ_0, with spin factors added and coupled to give spin eigenvalues (s, m). The factor Φ_0 is then defined in such a way that (II.2) holds; it will appear that Φ_0 need not be an eigenfunction of total spin and it is therefore sufficient to specify the spin z component, which will be $S - m$. Thus, with a more complete notation, (II.2) becomes

$$\Psi_0(1, 2, \ldots, N) = M'(\mu_{sm}^0)\mathsf{A}[\phi_0(\mu_{sm}|1, 2)\Phi_0(\mu_{S-m}|3, \ldots, N)] \qquad \text{(II.3)}$$

for every pair μ. This factorization is valuable because it preserves the separation into "pair" and "sea" functions, different pairs being equivalent merely to different states of a single two-electron group. To introduce correlation effects it is then necessary only to introduce "excited" pair functions (possibly containing the r_{12} variable). First, however, we note that the above product reduces to Ψ_0 on antisymmetrizing only because each factor is itself an antisymmetrized *orbital* product. In general, to obtain a function with spin eigenvalues S, M $(M = S)$, it will be necessary to add vector coupling coefficients and sum over a number of products corresponding to suitable choices of m. On replacing the pair function $\phi_0(\mu_{sm}|1, 2)$ by an *excited* pair function $\phi(\mu_{sm}|1, 2)$, and suitably coupling the spins, we therefore obtain a μ-excited function with the same spin eigenvalues as Ψ_0:

$$\Psi_\mu(1, 2, \ldots, N) = \sum_m C_m \Psi(\mu_{sm}, \mu_{S-m}^0|1, 2, \ldots, N) \qquad \text{(II.4a)}$$

where

$$\Psi(\mu_{sm}, \mu_{S-m}^0|1, 2, \ldots, N) = M'(\mu_{sm})\mathsf{A}[\phi(\mu_{sm}|1, 2)\Phi_0(\mu_{S-m}|3, \ldots, N)]. \qquad \text{(II.4b)}$$

In both (II.3) and (II.4) attention is focused on a particular pair of electrons, those which occupy orbitals R, S $(=\mu)$.

The detailed form of the spin-coupled functions must now be discussed. The identities involved in writing (II.1) in the form (II.3) may be derived easily, by writing the antisymmetrizer[2]

$$\mathsf{A} = \frac{1}{N!}\sum_{\mathsf{P}} (-1)^p \mathsf{P}$$

---

[2] A is defined so as to have the projection operator property $\mathsf{A}^2 = \mathsf{A}$.

in terms of *partial* antisymmetrizers for complementary groups of electrons (A, B). Thus

$$\mathsf{A} = \frac{N_A!N_B!}{N!}\,\mathsf{A}'[\mathsf{A}_A\mathsf{A}_B]$$

where

$$\mathsf{A}' = \sum_{\mathsf{T}} (-1)^t\mathsf{T}$$

is a sum of transpositions *between* groups A and B, and where N_A and N_B are the numbers of electrons in groups A and B. If now F is a spin-orbital product (all spin-orbitals different), with complementary factors F_A, F_B, then

$$\Psi_A = \sqrt{N_A!}\,\mathsf{A}_A F_A, \qquad \Psi_B = \sqrt{N_B!}\,\mathsf{A}_B F_B$$

are normalized antisymmetrical functions for groups A and B, while

$$\Psi = \sqrt{N!}\,\mathsf{A}[F_A F_B] = \sqrt{\frac{N!}{N_A!N_B!}}\,\mathsf{A}[\Psi_A\Psi_B] = \sqrt{\frac{N_A!N_B!}{N!}}\,\mathsf{A}'[\Psi_A\Psi_B]$$

is normalized and antisymmetrical for the whole system. These results allow us to write Ψ_0, defined in (II.3), in the more convenient form

$$\begin{aligned}\Psi_0 &= \Psi(\mu_{sm}^0, \mu_{S-m}^0|1, \ldots, N)\\ &= M(\mu_{sm}^0)\sqrt{\frac{2}{N(N-1)}}\,\mathsf{A}'[\phi_0(\mu_{sm}|1, 2)\Phi_0(\mu_{S-m}|3, \ldots, N)].\end{aligned} \tag{II.5}$$

The possible choices of the "pair" and "sea" factors are listed in Table I; it should be noted that this factorization of Ψ_0 is accomplished for *any* pair and for any m value.

The excited functions in (II.4b) may be written in a form analogous to (II.5):

$$\begin{aligned}\Psi(\mu_{sm}, \mu_{S-m}^0|1, \ldots, N) &= M(\mu_{sm})\sqrt{\frac{2}{N(N-1)}}\\ &\quad\times \mathsf{A}'[\phi(\mu_{sm}|1, 2)\Phi_0(\mu_{S-m}|3, \ldots, N)].\end{aligned} \tag{II.6}$$

They differ from those in (II.5) only in that $\phi(\mu_{sm})$ is in general a *correlated* pair function and, representing essentially a small correction, is not normalized to unity. It is most convenient to fix this normalization so that

$$M(\mu_{sm})M(\mu_{sm}^0) = 1 \tag{II.7}$$

where the $M(\mu_{sm}^0)$ are the factors in Table I, since this choice will simplify the matrix element expressions involving Ψ_0 and Ψ_μ. It is also readily

TABLE I

PAIR-SEA FACTORIZATION OF Ψ_0[a]

Pair	s	m	Normalizing factor $M(\mu^0_{sm})$	Pair function $\phi_0(\mu_{sm}\|1, 2)$	Sea function $\Phi_0(\mu_{S-m}\|3 \ldots N)$
$\mu = (AA)$	0	0	1	$\det(\psi_A\bar{\psi}_A)$	$\det(\psi_B\bar{\psi}_B \ldots \psi_X)$
$\mu = (AB)$	0	0	1	$(2)^{-1/2}[\det(\psi_A\bar{\psi}_B) - \det(\bar{\psi}_A\psi_B)]$	$(2)^{-1/2}[\det(\bar{\psi}_A\psi_B\psi_C \ldots \psi_X)$ $-\det(\psi_A\bar{\psi}_B\psi_C \ldots \psi_X)]$
	1	1		$\det(\psi_A\psi_B)$	$\det(\bar{\psi}_A\bar{\psi}_B\psi_C \ldots \psi_X)$
	1	0	$(-1)^m$	$(2)^{-1/2}[\det(\psi_A\bar{\psi}_B) + \det(\bar{\psi}_A\psi_B)]$	$(2)^{-1/2}[\det(\bar{\psi}_A\psi_B\psi_C \ldots \psi_X)$ $+\det(\psi_A\bar{\psi}_B\psi_C \ldots \psi_X)]$
	1	-1		$\det(\bar{\psi}_A\bar{\psi}_B)$	$\det(\psi_A\psi_B\psi_C \ldots \psi_X)$
$\mu = (AU)$	0	0	$\sqrt{2}(-1)^{\nu_A+\nu_U-1}$	$(2)^{-1/2}[\det(\psi_A\bar{\psi}_U) - \det(\bar{\psi}_A\psi_U)]$	$\det(\psi_A\psi_B\bar{\psi}_B \ldots \psi_V \ldots \psi_X)$
	1	1	$(-1)^{\nu_A+\nu_U-1}$	$\det(\psi_A\psi_U)$	$\det(\bar{\psi}_A\psi_B\bar{\psi}_B \ldots \psi_V \ldots \psi_X)$
	1	0	$\sqrt{2}(-1)^{\nu_A+\nu_U}$	$(2)^{-1/2}[\det(\psi_A\bar{\psi}_U) + \det(\bar{\psi}_A\psi_U)]$	$\det(\psi_A\psi_B\bar{\psi}_B \ldots \psi_V \ldots \psi_X)$
	1	-1	(not required in coupling to ground function multiplicity)		
$\mu = (UV)$	1	1	$(-1)^{\nu_U+\nu_V-1}$ $(\nu_U < \nu_V)$	$\det(\psi_U\psi_V)$	$\det(\psi_A\bar{\psi}_A \ldots \psi_W \ldots \psi_X)$

[a] Each factor of type $\det(\psi_R\psi_S \ldots)$ is a *normalized* antisymmetrized orbital product. The table gives the factors in the representation of Ψ_0 given by Eq. (II.5); every choice leads identically to the spin eigenfunction Ψ_0. In the phase factors, ν_R is the serial number of ψ_R in the spin-orbital sequence appearing in Ψ_0. Although ψ_A is shown first ($\nu_A = 1$) the results remain true when A, B are the names of any closed-shell orbitals and U, V those of any open-shell orbitals.

shown by straightforward spin algebra that, when recoupling of spins is confined to the orbitals left singly occupied by a pair excitation, the spin eigenfunctions (II.4a) take a very simple form in which [with the normalization (II.7)] all nonzero coefficients are equal. When the excited pairs are coupled to spin eigenvalues s, m, the μ-excited functions with spin eigenvalues S, M $(M = S)$ are found to be

$$\Psi_\mu = \sum_m \Psi(\mu_{sm}, \mu^0_{S-m}|1, \ldots, N) \qquad \text{(II.8)}$$

where the component functions are obtained on replacing $\phi_0(\mu_{sm})$ in Table I by the correlated pair function $\phi(\mu_{sm})$ defined in Table II.

TABLE II

The Correlated Pair Functions[a]

s	m	Pair function $\phi(\mu_{sm}\|1, 2)$	Symmetry property
0	0	$f_\mu(1, 2)\frac{1}{\sqrt{2}}(\alpha\beta - \beta\alpha)$	$f_\mu(1, 2) = +f_\mu(2, 1)$
1	1	$f_\mu(1, 2)\alpha\alpha$	$f_\mu(1, 2) = -f_\mu(2, 1)$
1	0	$f_\mu(1, 2)\frac{1}{\sqrt{2}}(\alpha\beta + \beta\alpha)$	
1	-1	$f_\mu(1, 2)\beta\beta$	

[a] The table gives the space and spin factors which make up the pair functions $\phi(\mu_{sm})$.

The actual form of the excited pair functions will be considered only briefly at this point. If the $\phi(\mu_{sm})$ in (II.6) were constructed simply from orbital products, drawn from the orthogonal complement of the set appearing in (II.1), a large number of excitations would be necessary in order to represent adequately the correlation of the two electrons; this would in fact correspond to the straightforward method of configuration interaction, which is known to be very slowly convergent. There are two methods of meeting this situation. The first is in essence that of Brueckner (1955), Goldstone (1957), and others [for a simple review, see Rodberg (1957)], in which all excitations of a given type are admitted and their effects are summed by diagram techniques; the second is that of Sinanoğlu (1961, 1962, 1964) and Szasz (1959, 1960, 1962a,b, 1963) in which a *single* excited function is admitted for each pair, but with a much more general functional form which may be optimized in order to obtain a highly accurate description of the correlation effects. This second approach, with which the present paper is concerned, leads to a variational equation which completely determines $\phi(\mu_{sm})$ for every pair.

In order to admit the excited functions it is necessary to obtain matrix elements of the full Hamiltonian H(1 ... N) between any two generalized products. This problem has been solved elsewhere (McWeeny, 1959, 1960) for group functions which are "strong orthogonal" in the sense

$$\int \phi_0^*(\mu|1, i)\Phi_0(\mu|1, j, k, \ldots, N)\, dx_1 = 0 \tag{II.9}$$

and even for functions in which no kind of orthogonality is assumed (McWeeny and Sutcliffe, 1963). It is, of course, vastly simpler to treat the orthogonal case and it is therefore of special interest to note (Sinanoğlu, 1961, 1962, 1964; Szasz, 1959, 1960, 1962a,b, 1963) that, *provided* $\Phi_0(\mu)$ *is an orbital product* (i.e., belongs to a finite subspace of the many-electron Hilbert space), strong orthogonality can be assumed without loss of generality. To see this it is only necessary to introduce the projection operator $\rho_0^{(\mu)}$ for the subspace spanned by the orbitals which appear in $\Phi_0(\mu)$, and the corresponding annihilator $1 - \rho_0^{(\mu)} = \omega_0^{(\mu)}$. Explicitly, $\rho_0^{(\mu)}$ is an integral operator[3] with kernel

$$\rho_0^{(\mu)}(x; x') = \sum_{[R \text{ in } \Phi_0(\mu)]} \psi_R(x)\psi_R^*(x'). \tag{II.10}$$

Thus the projection operator $\rho_0^{(\mu)}(1)$ in the representation space of particle 1 projects out that part of an arbitrary function $f(1, \ldots, n)$ which can be expanded in terms of the ψ_R in $\Phi_0(\mu)$ but destroys the remainder. On the other hand, $\omega_0^{(\mu)}(1)$ projects out the part which can be expanded only over the orthogonal complement of the set in $\Phi_0(\mu)$; consequently, the scalar product

$$\int \Phi_0^*(\mu|1, 2, \ldots, N)\omega_0^{(\mu)}(1) f(1, 2, \ldots, n)\, dx_1$$

must vanish. In order that a group be strong orthogonal for integration over two variables it is necessary to use both annihilators; then

$$\bar{\phi}(\mu|1, 2) = \omega_0^{(\mu)}(1)\omega_0^{(\mu)}(2)\phi(\mu|1, 2) \tag{II.11}$$

---

[3] Note that this operator has a convenient symbolic form. Operating on any function ψ, $\psi_R\psi_R{}^*\psi = \psi_R\langle\psi_R|\psi\rangle = \text{constant} \times \psi_R$; similarly $\Sigma_R\,\psi_R\psi_R{}^*$ projects onto a whole subspace. Equation (II.10) merely defines such an operator in the Schrödinger representation; its effect is then $\rho_0{}^{(\mu)}\psi(x) = \int \rho_0{}^{(\mu)}(x; x')\psi(x')\, dx'$.

is strong orthogonal to $\Phi_0(\mu)$ for any choice of $\phi(\mu)$ [trivial for $\phi_0(\mu)$]. It is also readily seen, as Szasz has remarked, that replacement of $\phi(\mu)$ in $\Psi(\mu)$ by $\tilde{\phi}(\mu)$ leaves $\Psi(\mu)$ unchanged. Thus, when one group function is "correlated" (in general containing interelectronic distances) but the other is built from a finite number of orbital products, the assumption of strong orthogonality between them implies no loss of generality. On the other hand, it may often be convenient to replace $\phi(\mu)$ by the function

$$\tilde{\phi}(\mu|1, 2) = \omega_0(1)\omega_0(2)\phi(\mu|1, 2) \tag{II.12}$$

where

$$\omega_0(i) = 1 - \rho_0(i) \tag{II.13}$$

and $\rho_0(i)$ is the projection operator onto the *entire* subspace of the spin-orbitals in Ψ_0, with kernel

$$\rho_0(x_i; x_i') = \sum_{(R \text{ in } \Psi_0)} \psi_R(x_i)\psi_R^*(x_i'). \tag{II.14}$$

In this case, as Szasz has pointed out, a constraint is implied which, in orbital language, is equivalent to the omission of configurations with a single orbital excitation. In many cases this constraint may be expected to be weak (cf. Brillouin's theorem) but when $S \neq 0$ the singly excited functions may be of more interest. They describe "polarization" effects, however, rather than correlation, and discussion is therefore deferred to a later section. Here we are concerned with the excited *pair* functions, which we now take to be of type (II.12), strong orthogonal to all the orbitals in Ψ_0, though in future we suppress the tilde.

The matrix elements of the Hamiltonian between two excited functions may be obtained by noting that the Ψ_μ are built up from "group products" which have the general form [cf. (II.6)]

$$\begin{aligned}\Psi_\kappa(1 \cdots N) &= \Psi_{Aa,Bb}(1 \cdots N)\\ &= \sqrt{\frac{N_A!N_B!}{N!}}\, \mathrm{A}'[\Phi_{Aa}(1 \cdots N_A)\Phi_{Bb}(\overline{N_A + 1} \cdots \overline{N_A + N_B})]\end{aligned} \tag{II.15}$$

where the "A group" consists of N_A electrons in state a, and the "B group" consists of the remaining N_B electrons in state b, with $N_A + N_B = N$. The single subscript κ indicates this "configuration," $\kappa = (Aa, Bb)$. Matrix element expressions between functions Ψ_κ, $\Psi_{\kappa'}$ of this kind, with strong orthogonal factors, have been given elsewhere (McWeeny, 1959, 1960). They require slight modification if the functions are not normalized, as in the case of Ψ_μ, and then take the following form:

Overlap Integral:

$$\langle\Psi_{\kappa'}|\Psi_\kappa\rangle = \langle Aa'|Aa\rangle\langle Bb'|Bb\rangle \tag{II.16}$$

where κ' indicates group states Aa', Bb', and the factors on the right are overlap integrals for the functions of the individual groups.

One-Electron Density Matrix:

$$\rho_1(\kappa\kappa'|1;1') = \rho_1^A(aa'|1;1')\langle Bb|Bb'\rangle + \langle Aa|Aa'\rangle\rho_1^B(bb'|1;1') \tag{II.17}$$

It should be noted that when Ψ_κ is of the form (II.6) the density matrix $\rho_1(0\kappa\,|\,1;\,1')$ connecting Ψ_κ and Ψ_0 will vanish because of strong orthogonality of the pair function to all the orbitals in Ψ_0; this introduces further simplifications.

Two-Electron Density Matrix:

$$\begin{aligned}\rho_2(\kappa\kappa'|1,2;1',2') = {} & \rho_2^A(aa'|1,2;1',2')\langle Bb|Bb'\rangle \\ & + \langle Aa|Aa'\rangle\rho_2^B(bb'|1,2;1',2') \\ & + \rho_1^A(aa'|1;1')\rho_1^B(bb'|2;2') \\ & - \rho_1^A(aa'|2;1')\rho_1^B(bb'|1;2') \\ & + \rho_1^A(aa'|2;2')\rho_1^B(bb'|1;1') \\ & - \rho_1^A(aa'|1;2')\rho_1^B(bb'|2;1')\end{aligned} \tag{II.18}$$

When Ψ_κ is of the form (II.6) the density matrix $\rho_2(0\kappa|1,2;1',2')$ given by (II.18) reduces to the first term owing to the strong orthogonality.

III. The Calculation of Ψ_0: SCF-HF Method

Consider the one-determinant approximation (II.1), namely

$$\Psi_0(1\cdots N) = M_0\mathsf{A}[\psi_A(1)\bar{\psi}_A(2)\cdots\psi_U(i)\cdots\psi_X(N)]. \tag{III.1}$$

With the subspace of the occupied spin orbitals is associated the projection operator defined in (II.14)

$$\rho_0 = \sum_{(R\text{ in }\Psi_0)} \psi_R\psi_R^* \tag{III.2}$$

which in the present case reduces to

$$\begin{aligned}\rho_0 &= \left(\sum_A AA^* + \sum_U UU^*\right)\alpha\alpha^* + \left(\sum_A AA^*\right)\beta\beta^* \\ &= (\mathsf{R}_1 + \mathsf{R}_2)\alpha\alpha^* + \mathsf{R}_1\beta\beta^* \qquad \text{(III.3)}\end{aligned}$$

where $\mathsf{R}_1 = \sum_A AA^*$ and $\mathsf{R}_2 = \sum_U UU^*$. It is convenient to use these symbolic forms of the projection operators in what follows. In a discrete representation the corresponding operator equations turn into matrix equations, which have been discussed elsewhere (McWeeny, 1964), while in a continuous representation they become integral equations. The "trace" of an operator will be interpreted as the trace of its representative matrix $(\sum_r \rho_{rr})$ or the corresponding integral $[\int \rho(x;\, x)\,dx]$, this being invariant against change of representation. The symbolic form is advantageous in what follows, but the results will be stated only briefly since the parallel matrix derivation is available elsewhere (McWeeny, 1964).

The spinless density operators R_1, R_2 satisfy the conditions

$$\mathsf{R}_1^2 = \mathsf{R}_1, \qquad \mathsf{R}_2^2 = \mathsf{R}_2, \qquad \mathsf{R}_1\mathsf{R}_2 = \mathsf{R}_2\mathsf{R}_1 = 0. \qquad \text{(III.4)}$$

We define coulomb and exchange operators in the usual way:

$$\begin{aligned}\mathsf{J}(1) &= \int g(1,2)\rho_0(2;2)\,dx_2 \\ \mathsf{K}(1) &= \int dx_2\, g(1,2)\rho_0(1;2)\mathsf{P}(1\to 2)\end{aligned} \qquad \text{(III.5)}$$

where $\mathsf{P}(1 \to 2)$ changes variable 1 to 2, and find

$$E = \operatorname{tr}(\mathsf{f} + \tfrac{1}{2}\mathsf{G})\rho_0 \qquad \text{(III.6)}$$

where $\mathsf{G} = \mathsf{J} - \mathsf{K}$. By considering a first-order variation in ρ_0, the first-order change in E is found to be

$$\delta E = \operatorname{tr}\, \mathsf{h}^F \delta\rho_0 \qquad \text{(III.7)}$$

where

$$\mathsf{h}^F = \mathsf{f} + \mathsf{G} = \mathsf{f} + \mathsf{J} - \mathsf{K}. \qquad \text{(III.8)}$$

If we define the spinless operators

$$\mathsf{h}^F_{\alpha\alpha} = \alpha^*\mathsf{h}^F\alpha, \qquad \mathsf{h}^F_{\beta\beta} = \beta^*\mathsf{h}^F\beta, \qquad \text{(III.9)}$$

then

$$\delta E = \operatorname{tr}\{\mathsf{h}^F_{\alpha\alpha}(\delta\mathsf{R}_1 + \delta\mathsf{R}_2) + \mathsf{h}^F_{\beta\beta}\delta\mathsf{R}_1\}. \qquad \text{(III.10)}$$

The variation is subject to the conditions (III.4). Here R_1 and R_2 are projection operators which define subspaces $\mathbf{S}_1$ and $\mathbf{S}_2$; and therefore $R_3 = (1 - R_1 - R_2)$ defines the complementary space $\mathbf{S}_3$. Using the projection operator methods discussed elsewhere (McWeeny, 1962, 1964), the following stationary value conditions are obtained:

$$\begin{aligned} R_3 h_1 R_1 &= 0, & h_1 &= \tfrac{1}{2}(h^F_{\alpha\alpha} + h^F_{\beta\beta}), \\ R_3 h_2 R_2 &= 0, & h_2 &= h^F_{\alpha\alpha}, \\ R_1 D R_2 &= 0, & D &= h^F_{\beta\beta} = 2h_1 - h_2. \end{aligned} \tag{III.11}$$

It is interesting to note that these conditions are satisfied when the orbitals themselves are solutions of the pseudoeigenvalue equations (well-known in open-shell SCF theory):

$$\begin{aligned} (1 - R_2) h_1 A &= \varepsilon_A A, \\ (1 - R_1) h_2 U &= \varepsilon_U U. \end{aligned} \tag{III.12}$$

In more symmetrical form

$$\begin{aligned} \bar{h}_1 A &= \varepsilon_A A, & \bar{h}_1 &= (1 - R_2) h_1 (1 - R_2), \\ \bar{h}_2 U &= \varepsilon_U U, & \bar{h}_2 &= (1 - R_1) h_2 (1 - R_1), \end{aligned} \tag{III.13}$$

where h_1 and $\bar{h}_2$ are effective Hamiltonians for closed and open shells, respectively. It is more convenient, however, to define a single Hamiltonian, equivalent to that introduced by Roothaan (1960), which has *both* open- and closed-shell orbitals as its eigenfunctions. This may be derived by noting that the necessary and sufficient condition[4] for (III.11) to be satisfied is that the orbitals R $(= A, \ldots, U, \ldots)$ satisfy

$$\bar{h} R = \varepsilon_R R \tag{III.14}$$

where

$$\bar{h} = a\bar{h}_1 + b\bar{h}_2 + c\bar{h}_3 \tag{III.15}$$

and

$$\bar{h}_3 = (1 - R_3) D (1 - R_3) = (R_1 + R_2) D (R_1 + R_2).$$

The parameters a, b, and c are arbitrary insofar as $\bar{h}$ always has the same set of eigenfunctions, but it is most convenient to define the independent-particle "model" in terms of

$$h^M = \tfrac{1}{2}[\bar{h}_1 + \bar{h}_2 + \bar{h}_3]. \tag{III.16}$$

---

[4] To within a unitary transformation.

It is also clear that the *spin* orbitals ψ_R (obtained by adding spin factor α or β) all satisfy the same eigenvalue equation; in fact

$$\mathsf{h}^M\psi_R = \varepsilon_R\psi_R \qquad (\text{all } \psi_R). \tag{III.17}$$

The one-determinant approximation Ψ_0 is therefore an eigenfunction of the "model" Hamiltonian

$$\mathsf{H}_0 = \sum_{i=1}^{N} \mathsf{h}^M(i) \tag{III.18}$$

with eigenvalue

$$E_0 = \sum_R \varepsilon_R = 2\sum_A \varepsilon_A + \sum_U \varepsilon_U. \tag{III.19}$$

The essence of Sinanoğlu-type theories is their use of a model Hamiltonian H_0 as the zeroth-order Hamiltonian in a perturbation variation treatment. In the open-shell case it seems to be particularly important to use the Hamiltonian (III.18) which recognizes spin eigenvalue constraints and at the same time gives the best one-determinant energy. So far, this formulation does not appear to have been utilized; it will be developed in the next section.

IV. Perturbation Variation Method

Consider the wave function

$$\Psi = \Psi_0 + X. \tag{IV.1}$$

The model (zeroth-order) Hamiltonian may be written

$$\mathsf{H}_0(1 \cdots N) = \sum_{i=1}^{N} \mathsf{h}^M(i) = \sum_{i=1}^{N} [\mathsf{f}(i) + \mathsf{G}^M(i)] \tag{IV.2}$$

where $\mathsf{G}^M(i)$ is a complicated "average interaction" defined by reference to (III.16). The complete (nonrelativistic) Hamiltonian is then

$$\mathsf{H}(1 \cdots N) = \sum_i \mathsf{f}(i) + \tfrac{1}{2}\sum_{i \neq j} \mathsf{g}(i,j) = \mathsf{H}_0 + \mathsf{H}_1 \tag{IV.3}$$

where

$$\mathsf{H}_1 = \tfrac{1}{2}\sum_{i \neq j} \bar{\mathsf{g}}(i,j) \tag{IV.4a}$$

and

$$\bar{\mathsf{g}}(i,j) = \mathsf{g}(i,j) - \frac{\mathsf{G}^M(i) + \mathsf{G}^M(j)}{N-1}. \tag{IV.4b}$$

Since $\bar{\mathsf{g}}$ describes the difference between the "instantaneous" potential g and the "average" used in the Hartree–Fock model it may be referred to

(Sinanoğlu, 1961, 1962, 1964) as the "fluctuation potential," and should represent a relatively weak interaction. We assume first that Ψ_0 is an *exact* eigenfunction of H_0, so that the first-order correction X may be determined by standard perturbation theory. Then, to first order,

$$(\mathsf{H}_0 - E_0)X = -(\mathsf{H}_1 - E_1)\Psi_0, \tag{IV.5}$$

and the first-order function X gives the energy to third order:

$$\begin{aligned} E_1 &= \langle\Psi_0|\mathsf{H}_1|\Psi_0\rangle, \\ E_2 &= \langle\Psi_0|\mathsf{H}_1|X\rangle - E_1\langle\Psi_0|X\rangle, \\ E_3 &= \langle X|\mathsf{H}_1|X\rangle - E_1\langle X|X\rangle - 2E_2\langle\Psi_0|X\rangle. \end{aligned} \tag{IV.6}$$

Thus

$$\begin{aligned} E &= E_0 + E_1 + E_2 + E_3 + \cdots \\ &= E_{\text{SCF}} + E_2 + E_3 + \cdots \end{aligned} \tag{IV.7}$$

where we note the well-known fact that E_1 exactly compensates for the double counting of electron interactions in the orbital energy sum (E_0) to give the usual SCF expression. Unfortunately (IV.5) is as difficult to solve as the original problem. Also, approximate solution by expansion over unperturbed eigenfunctions leads to difficulties connected with convergence and inclusion of the continuum. For these reasons it is preferable to introduce

$$X = \sum_\mu \Psi_\mu \tag{IV.8}$$

where Ψ_μ is of the form (II.8) and then use a variational procedure to determine the optimum Ψ_μ. The variational equivalent of (IV.5) is

$$E_2 \leqq W_2 = \langle X|\mathsf{H}_0 - E_0|X\rangle + 2\langle X|\mathsf{H}_1 - E_1|\Psi_0\rangle. \tag{IV.9}$$

If X is expanded in the form (IV.8), where the functions Ψ_μ involve excited pair functions of the form (II.11), there will occur cross-terms containing excited functions corresponding to two different pairs. If, however, we use (II.12) and the fact that the orbitals satisfy (III.17), it follows easily that such cross-terms disappear. Equation (IV.9) can then be written

$$W_2 = \sum_\mu W_2'(\mu) \tag{IV.10}$$

where

$$W_2'(\mu) = \langle\Psi_\mu|\mathsf{H}_0 - E_0|\Psi_\mu\rangle + 2\langle\Psi_\mu|\mathsf{H}_1 - E_1|\Psi_0\rangle. \tag{IV.11}$$

The strong orthogonality of the chosen pair functions [see (II.12) *et seq.*], therefore, leads to a *separation* of equations for the different excited pair functions, equations which would otherwise be coupled. The stationary condition now becomes

$$W_2'(\mu) = \text{stationary value} \tag{IV.12}$$

for each excited pair.

The matrix elements in (IV.11) are easily reduced for excited functions of the form (II.6) and (II.8). From (II.17), and the fact that the sea functions Φ_0 are normalized, it follows that

$$\begin{aligned}\langle\Psi_\mu|\mathsf{H}_0 - E_0|\Psi_\mu\rangle = \sum_m M(\mu_{sm})^2 \Big[&\mathrm{tr}\ \mathsf{h}^M(1)\rho_1^{(\mathrm{pair})}(\mu_{sm}\mu_{sm}|1;1') \\ &+ \langle\phi(\mu_{sm})|\phi(\mu_{sm})\rangle[\mathrm{tr}\ \mathsf{h}^M(1)\rho_1^{(\mathrm{sea})}(\mu_{S-m}\mu_{S-m}|1;1') \\ &- \sum_R \varepsilon_R]\Big].\end{aligned}$$

This may be written, introducing $\varepsilon_\mu = \varepsilon_R + \varepsilon_S$ for $\mu = (RS)$,

$$\langle\Psi_\mu|\mathsf{H}_0 - E_0|\Psi_\mu\rangle = D_\mu\langle\phi(\mu_{sm}|1,2)|\mathsf{h}^M(1) + \mathsf{h}^M(2) - \varepsilon_\mu|\phi(\mu_{sm}|1,2)\rangle \tag{IV.13}$$

where

$$D_\mu = \sum_m M(\mu_{sm})^2. \tag{IV.14}$$

The integral in (IV.13) is independent of the value of m, which runs from $-s$ to $+s$. The remaining matrix element in (IV.11) reduces, owing to strong orthogonality between $\phi(\mu_{sm})$ and Ψ_0 which makes the one-electron density matrix (II.17) vanish, to a two-electron term:

$$\langle\Psi_\mu|\mathsf{H}_1 - E|\Psi_0\rangle = \tfrac{1}{2}\ \mathrm{tr}\ \bar{\mathsf{g}}(1,2)\rho_2(0\mu|1,2;1',2'). \tag{IV.15}$$

Since Ψ_μ is a sum of terms (II.8) the transition density matrix reduces; we use (II.7) and write (IV.15) in the alternative form

$$\langle\Psi_\mu|\mathsf{H}_1 - E_1|\Psi_0\rangle = D_\mu\langle\phi(\mu_{sm}|1,2)|\bar{\mathsf{g}}(1,2)|\phi_0(\mu_{sm}|1,2)\rangle. \tag{IV.16}$$

Finally, then, the approximate second-order energy (IV.10) takes the form

$$W_2 = \sum_\mu D_\mu W_2(\mu) \tag{IV.17}$$

where the weight factors D_μ are defined in (IV.14), and optimum pair functions follow from

$$\begin{aligned}W_2(\mu) &= \langle\phi(\mu_{sm}|1,2)|\mathsf{h}^M(1) + \mathsf{h}^M(2) - \varepsilon_\mu|\phi(\mu_{sm}|1,2)\rangle \\ &\quad + 2\langle\phi(\mu_{sm}|1,2)|\bar{\mathsf{g}}(1,2)|\phi_0(\mu_{sm}|1,2)\rangle \\ &= \text{stationary value.}\end{aligned} \tag{IV.18}$$

This variational equation may of course be solved for any convenient m value, $|m| \leqq s$. For a pair, $s = 0$ or 1, and the values of D_μ are listed in Table III for the various types of excitation.

TABLE III

WEIGHT FACTORS IN THE SECOND-ORDER ENERGY

μ	D_μ (singlet)	D_μ (triplet)
AA	1	0
AB	1	3
AU	$\frac{1}{2}$	$\frac{3}{2}$
UV	0	1

The third-order energy E_3 will not be considered in the present paper. In general, the energy expressions become rather more complicated, involving cross-terms between all the excited pair functions. These terms are simple only for AA-type excitations in a closed-shell system in which case they have been discussed by Krauss and Weiss (1964).

The N-electron problem has now been replaced by a set of two-electron problems involving the correlated pair functions. One or two remarks on their meaning and on methods of solution are necessary. First, we note that a direct variational solution of (IV.18) may be regarded as an approximate solution of the *two*-body version of (IV.5), namely,

$$(\mathsf{H}_0 - E_0^{(\mu)})\phi(\mu_{sm}|1, 2) = -(\mathsf{H}_1 - E_1^{(\mu)})\phi_0(\mu_{sm}|1, 2) \qquad \text{(IV.19)}$$

in which

$$\mathsf{H}_0 = \mathsf{h}^M(1) + \mathsf{h}^M(2), \qquad \mathsf{H}_1 = \bar{\mathsf{g}}(1, 2),$$

$$E_0^{(\mu)} = \varepsilon_\mu, \qquad E_1^{(\mu)} = \langle \phi_0(\mu_{sm}|1, 2)\bar{\mathsf{g}}(1, 2)|\phi_0(\mu_{sm}|1, 2)\rangle,$$

and $\phi_0(\mu_{sm}|1, 2)$ is an exact eigenfunction of H_0. Equation (IV.17) therefore appears to determine, to first order, the correlation correction for two electrons moving in the field defined by h^M, with the screened interaction $\bar{\mathsf{g}}$ (the fluctuation potential). This is not really the case, because severe constraints, whose form is not physically obvious, are implied by the strong orthogonality condition; with this reservation, however, (IV.19) provides a formal connection with a two-electron problem with Hamiltonian

$$\mathsf{h}^M(1) + \mathsf{h}^M(2) + \bar{\mathsf{g}}(1, 2)$$

and hence an intuitive basis for the pair-sea picture. Second, the strong orthogonality condition may be introduced explicitly into Eq. (IV.18) by introducing the annihilator ω_0 of (II.12). It is clear that ω_0 may be replaced by the (spinless) operator R_3 (Section III) which projects onto the subspace complementary to that spanned by the orbitals appearing in Ψ_0, and we therefore set

$$\phi(\mu_{sm}|1, 2) = \mathsf{R}_3(1)\mathsf{R}_3(2)\chi(\mu_{sm}|1, 2) \qquad \text{(IV.20)}$$

in which $\chi(\mu_{sm}|1, 2)$ may be varied *without* constraints.

It is then a simple matter to reduce the two-body equation (IV.18) by using the explicit forms of the model Hamiltonian and fluctuation potential. Thus $\bar{\mathsf{g}}$ in (IV.15) may be replaced by g, owing to the strong orthogonality of $\phi(\mu_{sm})$ and $\phi_0(\mu_{sm})$ which makes the one-electron density matrix (and hence the one-electron terms from $\bar{\mathsf{g}}$) vanish identically; and from (III.16) it follows that h^M may be replaced by $\frac{1}{2}(\mathsf{h}_1 + \mathsf{h}_2)$ as long as $\phi(\mu_{sm}|1, 2)$ has the projected form (IV.20). To summarize, with a variational function of the form (IV.20) we may make the replacements

$$\mathsf{h}^M \rightarrow \tfrac{1}{2}(\mathsf{h}_1 + \mathsf{h}_2), \qquad \bar{\mathsf{g}} \rightarrow \mathsf{g} \qquad \text{(IV.21)}$$

in Eq. (IV.18). Since these operators are spinless, the spin factors in (IV.18) may be removed at once. The operators h_1 and h_2 are well-known in open-shell SCF theory and are given below for completeness:

$$\begin{aligned} \mathsf{h}_1 &= \mathsf{f} + (2\mathsf{j}_1 - \mathsf{k}_1) + (\mathsf{j}_2 - \tfrac{1}{2}\mathsf{k}_2), \\ \mathsf{h}_2 &= \mathsf{f} + (2\mathsf{j}_1 - \mathsf{k}_1) + (\mathsf{j}_2 - \mathsf{k}_2), \end{aligned} \qquad \text{(IV.22)}$$

where, in terms of the spinless density matrices (Section III) for closed- and open-shell orbitals ($i = 1, 2$),

$$\begin{aligned} \mathsf{j}_i(1) &= \int dr_2\, g(1, 2)R_i(2; 2) \\ \mathsf{k}_i(1) &= \int dr_2\, g(1, 2)R_i(1; 2)\mathsf{P}(1 \rightarrow 2) \end{aligned} \qquad \text{(IV.23)}$$

are spinless coulomb and exchange operators.

Finally, we note that the variational equation (IV.18) and its derivation through perturbation theory depends on a knowledge of the *exact* solutions of the Hartree–Fock model. Normally these are not available and it is important to discuss how the equations must be modified when *approximate* Hartree-Fock orbitals are employed, these being built up using a *truncated* one-electron basis. It will also be necessary, at this point, to investigate the effect of *one*-electron excitations, which will become important as a means of removing the residual error in the approximate Hartree–Fock description.

V. The General First-Order Wave Function

In this section we investigate in particular (*i*) the admission of single excitations and (*ii*) the effect of using an approximate, instead of exact, Hartree–Fock function Ψ_0. As a preliminary, it is necessary to have an energy expression which is formally exact and is not based upon a power series expansion in a perturbation parameter. To obtain such an expression we again write

$$\mathsf{H} = \mathsf{H}_0 + \mathsf{H}_1 \tag{V.1}$$

where H_0 is the model Hamiltonian (IV.2), and suppose that the *exact* wave function is

$$\Psi = \Psi_0 + X. \tag{V.2}$$

In this case X may be assumed to have the form (e.g., Sinanoğlu, 1964)

$$X = \sum_n X_n \tag{V.3}$$

where X_n is a sum of n-electron excited functions analogous to (IV.8). Each X_n is here strong orthogonal to Ψ_0 and we shall be particularly interested in the one- and two-electron terms ($n = 1, 2$). Simple manipulation of the expression $E = \langle \Psi|\mathsf{H}|\Psi\rangle$ then gives the exact result

$$E = E_0 + E_1 + E_2 + \frac{E_3 - E_2\langle X|X\rangle}{1 + \langle X|X\rangle} \tag{V.4}$$

where

$$\begin{aligned} E_0 &= \langle \Psi_0|\mathsf{H}_0|\Psi_0\rangle, \\ E_1 &= \langle \Psi_0|\mathsf{H}_1|\Psi_0\rangle, \\ E_2 &= \langle X|\mathsf{H}_0 - E_0|X\rangle + 2\langle X|\mathsf{H}_1|\Psi_0\rangle, \\ E_3 &= \langle X|\mathsf{H}_1 - E_1|X\rangle. \end{aligned} \tag{V.5}$$

It is not assumed that Ψ_0 is an *exact* Hartree–Fock function (i.e., an exact eigenfunction of H_0), but should this be the case then (V.4) yields the perturbation expansion (IV.6) and E_1, E_2, and E_3 become the first-, second-, and third-order terms in the sense of perturbation theory. We shall continue to refer to E_1 and E_2 as "first- and second-order terms," even when an exact Hartree–Fock function is not available; the last term in (V.4) will be called the "remainder term" and denoted by E_R. The actual size of the terms in (V.4) will, however, depend on how H and Ψ are partitioned; provided the partitioning makes E_R small, only E_2 will depend significantly upon X and a variational treatment will yield the optimum excited functions (one-electron, two-electron, etc.) with which to supplement the given ground function Ψ_0. The theory developed so far represents the special case in which X contains only pair excitations and the partitioning in (V.1) and (V.2) is that of conventional perturbation theory. We now return to the two basic problems.

A. One- and Many-Electron Excitations

With a wave function of the form (V.2–V.3) in which the n-excited functions are all strong orthogonal to Ψ_0, it readily follows that

$$\begin{aligned} \langle X_m | X_n \rangle &= \delta_{mn} \langle X_n | X_n \rangle, \\ \langle X_n | \Psi_0 \rangle &= 0. \end{aligned} \tag{V.6}$$

Even without assuming that H_0 is the model Hamiltonian of (IV.2) and that Ψ_0 is an exact eigenfunction, (V.4) then becomes

$$\begin{aligned} E = E_0 + E_1 &+ 2\langle X | \mathsf{H}_0 | \Psi_0 \rangle \\ &+ \langle X | \mathsf{H}_0 - E_0 | X \rangle + 2\langle X | \mathsf{H}_1 | \Psi_0 \rangle + E_R. \end{aligned} \tag{V.7}$$

The first two terms again give the energy obtained from the initial SCF calculation:

$$E_0 + E_1 = E_{\text{SCF}} = \langle \Psi_0 | \mathsf{H} | \Psi_0 \rangle. \tag{V.8}$$

To reduce the other terms we note first that H_0 is a sum of one-electron operators and hence

$$\langle X | \mathsf{H}_0 | \Psi_0 \rangle = \langle X_1 | \mathsf{H}_0 | \Psi_0 \rangle, \tag{V.9}$$

$$\langle X_m | \mathsf{H}_0 - E_0 | X_n \rangle = 0, \qquad \text{unless } m = n,\ n \pm 1. \tag{V.10}$$

Similarly, H_1 is a sum of one- and two-electron operators, and consequently

$$\langle X | \mathsf{H}_1 | \Psi_0 \rangle = \langle X_1 + X_2 | \mathsf{H}_1 | \Psi_0 \rangle, \tag{V.11}$$

triple and higher excitations giving no contributions. With the assumption (which we shall relax presently) of an exact SCF Ψ_0, (V.9) and (V.10) reduce further to give

$$\langle X | \mathsf{H}_0 | \Psi_0 \rangle = 0, \tag{V.12}$$

$$\langle X_i | \mathsf{H}_0 - E_0 | X_j \rangle = 0, \qquad i \neq j. \tag{V.13}$$

On inserting these results in (V.7) we obtain, for an exact SCF Ψ_0,

$$E = E_{\text{SCF}} + \sum_{n=1}^{N} E_2(X_n) + E_R \tag{V.14}$$

where

$$E_2(X_n) = \langle X_n | \mathsf{H}_0 - E_0 | X_n \rangle + 2\langle X_n | \mathsf{H}_1 | \Psi_0 \rangle. \tag{V.15}$$

Now for $n \geqq 3$, the last term in (V.15) vanishes; and since E_0 refers to the ground state $(\mathsf{H}_0 - E_0)$ must have an expectation value whose lower bound is zero—this being achieved when X_n vanishes. In a variational procedure based on the second-order term E_2, triple and higher excitations may therefore be excluded. The resultant energy expression is thus

$$E = E_{\text{SCF}} + E_2(X_1) + E_2(X_2) + E_R. \tag{V.16}$$

When Ψ_0 is an exact SCF wave function, therefore, the exact first-order correction involves only single and double excitations and the corresponding second-order energy terms may be obtained from *independent* variational calculations on $E_2(X_2)$ and $E_2(X_1)$. Previous sections have been concerned with the second of these terms, introducing the pair correlations, but the single-excitation corrections may clearly be added without difficulty. For a closed-shell system, of course, $E_2(X_1) = 0$ (Brillouin's theorem), but this is not so in the open-shell case.

B. Truncation Effects

When Ψ_0 is *not* an exact eigenfunction of H_0, as would be the case for a function constructed using a finite basis of one-electron functions, the reductions in (V.12) and (V.13) no longer apply. The basic equation (V.7) then takes the form

$$E = E_{\text{SCF}} + 2\langle X_1 | \mathsf{H}_0 | \Psi_0 \rangle + \sum_{n=1}^{N} E_2(X_n) + 2\sum_{n=2}^{N} \langle X_{n-1} | \mathsf{H}_0 | X_n \rangle + E_R \tag{V.17}$$

where $E_2(X_n)$ is again given by (V.15). The expression (V.17) contains an extra single-excitation term, $2\langle X_1 | \mathsf{H}_0 | \Psi_0 \rangle$, and the triple- and higher-excitation terms also appear.

To examine the connection between the perturbation treatment based on the exact Ψ_0 and one based on an *approximate* SCF function, which we now denote by $\tilde{\Psi}_0$, we write

$$\mathsf{H}_0 = \tilde{\mathsf{H}}_0 + \mathsf{V} \tag{V.18}$$

where $\tilde{\mathrm{H}}_0$ is the *projection* of the exact model Hamiltonian H_0 onto the subspace defined by the truncated one-electron basis from which the orbitals of $\tilde{\Psi}_0$ are constructed. Since H_0 is a sum of one-electron operators, $\mathrm{h}^M(i)$, this amounts merely to projecting each term onto the basic one-electron subspace. The operator V will then contain the projected part in the complementary subspace, together with coupling terms connecting the two [cf. McWeeny (1962)]. From this definition it follows that:

$$\tilde{\mathrm{H}}_0\tilde{\Psi}_0 = \tilde{E}_0\tilde{\Psi}_0 \tag{V.19}$$

where $\tilde{E}_0$ is an approximation to E_0. The perturbation theory may then be redeveloped with $\tilde{\mathrm{H}}_0$ as the unperturbed operator (exact eigenfunction $\tilde{\Psi}_0$) and the new partitioning $\mathrm{H} = \mathrm{H}_0 + \mathrm{H}_1 = \tilde{\mathrm{H}}_0 + (\mathrm{H}_1 + \mathrm{V})$. We therefore take

$$\mathrm{H} = \tilde{\mathrm{H}}_0 + \tilde{\mathrm{H}}_1 \tag{V.20}$$

where

$$\tilde{\mathrm{H}}_1 = \mathrm{H}_1 + \mathrm{V} \tag{V.21}$$

is the perturbation operator. Instead of (V.16) we then obtain

$$E = \tilde{E}_{\mathrm{SCF}} + \tilde{E}_2(\tilde{X}_1) + \tilde{E}_2(\tilde{X}_2) + \tilde{E}_R \tag{V.22}$$

where the approximate SCF energy is

$$\tilde{E}_{\mathrm{SCF}} = \tilde{E}_0 + \langle\tilde{\Psi}_0|\tilde{\mathrm{H}}_1|\tilde{\Psi}_0\rangle \tag{V.23}$$

and

$$\tilde{E}_2(\tilde{X}_n) = \langle\tilde{X}_n|\tilde{\mathrm{H}}_0 - \tilde{E}_0|\tilde{X}_n\rangle + 2\langle\tilde{X}_n|\tilde{\mathrm{H}}_1|\tilde{\Psi}_0\rangle. \tag{V.24}$$

Since, to second order, single and double excitations are uncoupled, the functions $\tilde{X}_1$ and $\tilde{X}_2$ may be optimized separately; thus, for the pair excitations, we should minimize $\tilde{E}_2(\tilde{X}_2)$. Now $\tilde{\mathrm{H}}_1 = \mathrm{H}_1 + \mathrm{V}$, and V contains only one-electron operators and will give zero contribution to the last term in (V.24) for $n = 2$. Also, $\tilde{\mathrm{H}}_0$ in the first term may be replaced by H_0 if we are prepared to disregard the third-order quantity $-\langle\tilde{X}_2|\mathrm{V}|\tilde{X}_2\rangle$—which could be transferred to the remainder term $\tilde{E}_R$. In a second-order theory, therefore, we may replace (V.24) for $n = 2$ by

$$\tilde{E}_2(\tilde{X}_2) = \langle\tilde{X}_2|\mathrm{H}_0 - \tilde{E}_0|\tilde{X}_2\rangle + 2\langle\tilde{X}_2|\mathrm{H}_1|\tilde{\Psi}_0\rangle. \tag{V.25}$$

Now this is unchanged to second order if we replace $\tilde{E}_0$ and $\tilde{\Psi}_0$ by E_0 and Ψ_0. To third order its minimization will give the same results as minimization of

$$E_2(\tilde{X}_2) = \langle\tilde{X}_2|\mathrm{H}_0 - E_0|\tilde{X}_2\rangle + 2\langle\tilde{X}_2|\mathrm{H}_1|\Psi_0\rangle. \tag{V.26}$$

But this is simply the exact second-order expression (V.15), for $n = 2$, except that the pair variation function has been called $\tilde{X}_2$ instead of X_2. This is an important result since it shows that, to second order, the *approximate* estimate of a pair correlation energy, based on minimization of (V.25), should agree with that which would be obtained using an exact Hartree–Fock function. The estimated pair correlation energy should therefore be largely independent of the size of the basis used in the approximate SCF calculation, a conclusion which appears to be substantiated by results so far available [see, for example, Linderberg (1960)].

Turning next to the single excitations, we note that when $n = 1$ the operator V *does* contribute to the final term and that the truncation effect is therefore operative in determining the admixture of single-excitation functions. On writing (V.24), for $n = 1$, in the form

$$\tilde{E}_2(\tilde{X}_1) = \langle \tilde{X}_1 | \tilde{\mathsf{H}}_0 - \tilde{E}_0 | \tilde{X}_1 \rangle + 2\langle \tilde{X}_1 | \mathsf{H}_1 | \tilde{\Psi}_0 \rangle + 2\langle \tilde{X}_1 | \mathsf{H}_0 - \tilde{\mathsf{H}}_0 | \tilde{\Psi}_0 \rangle \tag{V.27}$$

it is clear that the final term is responsible for that part of the second-order correction of the energy which arises because $\tilde{\Psi}_0$ is not an exact solution of the SCF equations. The middle term introduces the correction already encountered in the perturbation expression (V.15), for $n = 1$, based on the exact Ψ_0. From (V.19), $\tilde{\mathsf{H}}_0$ in the last term may be discarded. Again, in second order, we may replace $\tilde{\mathsf{H}}_0$ by H_0 in the first term; the result is:

$$\tilde{E}_2(\tilde{X}_1) = \langle \tilde{X}_1 | \mathsf{H}_0 - \tilde{E}_0 | \tilde{X}_1 \rangle + 2\langle \tilde{X}_1 | \mathsf{H}_1 | \tilde{\Psi}_0 \rangle + 2\langle \tilde{X}_1 | \mathsf{H}_0 | \tilde{\Psi}_0 \rangle. \tag{V.28}$$

The origin of the extra single-excitation term in (V.17) is now clear: this term must be added to $E_2(X_1)$ in order to obtain, using the approximate Ψ_0 (i.e., $\tilde{\Psi}_0$), the same results to second order as would follow using the exact Ψ_0. On the other hand, if we wish to evaluate the energy only to second order, the terms $E_2(X_n)$ $(n > 2)$ and the coupling terms $\langle X_{n-1} | \mathsf{H}_0 | X_n \rangle$ may be absorbed into the remainder E_R, for the latter are seen to be of third order on putting $\mathsf{H}_0 = \tilde{\mathsf{H}}_0 + \mathsf{V}$ and remembering that the X_n have no component in the domain of $\tilde{\mathsf{H}}_0$, while the former take their least values when $X_n = 0$ $(n > 2)$.

It is now clear that, with a remainder term containing quantities of third and higher order in the usual sense of perturbation theory, (V.17) may be written

$$E = E_{\mathrm{SCF}} + E_2(X_1) + 2\langle X_1 | \mathsf{H}_0 | \Psi_0 \rangle + E_2(X_2) + E_R \tag{V.29}$$

which differs from (V.16), derived assuming an exact SCF function, only in a modification of the single-excitation term. The optimization of the single- and double-excitation terms is thus achieved by solving

$$E_2'(X_1) = \langle X_1 | \mathsf{H}_0 - E_0 | X_1 \rangle + 2\langle X_1 | \mathsf{H}_1 | \Psi_0 \rangle + 2\langle X_1 | \mathsf{H}_0 | \Psi_0 \rangle$$
$$= \text{stationary value,} \qquad \text{(V.30)}$$

$$E_2(X_2) = \langle X_2 | \mathsf{H}_0 - E_0 | X_2 \rangle + 2\langle X_2 | \mathsf{H}_1 | \Psi_0 \rangle = \text{stationary value.} \qquad \text{(V.31)}$$

VI. Summary and Discussion

The main points of this analysis may be summarized as follows.

(*i*) The theory of pair-correlated wave functions (Sinanoğlu, 1961, 1962, 1964; Szasz, 1959, 1960, 1962a,b, 1963) may be derived easily from the group function approach (McWeeny, 1959, 1960), one electronic group being a pair of electrons and the other being the remaining (N-2) electrons. There is a nontrivial spin-coupling problem when it is required that the correlated wave function shall remain an exact eigenfunction of the total spin operators S^2 and S_z. This difficulty has been resolved in a simple way, applicable even in the general case of an "open-shell" system with a non-singlet ground state.

(*ii*) The basic single determinant wave function, Ψ_0, used as the zero-order approximation in the perturbation calculation, may be conveniently optimized without relaxing spin eigenfunction requirements by introducing a single model Hamiltonian containing suitable projection operators. The eigenfunctions of this operator are the exact SCF functions for both the closed and open shells.

(*iii*) On exciting electrons two at a time from the basic determinant to correlated pair functions which are strong orthogonal to Ψ_0, the variational equation for the second-order energy separates into a set of *two*-electron equations, one for each pair function. The fact that these equations are uncoupled is a result of the strong-orthogonality condition—a constraint which is equivalent to disregarding a possible contribution from *single* excitations.

(*iv*) The effects of (*a*) admitting single and multiple excitations and (*b*) using a basic function Ψ_0 which is constructed from a finite orbital basis, and is therefore not exact, have been discussed in some detail. To second order of perturbation theory only the single and double excitations participate; the double excitations, introducing correlated pair functions, account for the bulk of the correlation energy and their admission should give results which are substantially independent of whether Ψ_0 is exact or approximate. The equations for the singly excited functions are independent of those for the pair functions; they may be solved separately and give an additional energy term which corrects, to second order, for any defects in the initial SCF function Ψ_0. Truncation of the basis set in the SCF calculation is therefore not a serious obstacle to progress, and estimates of correlation energy based on approximate SCF functions should still be meaningful.

There are of course many other points of interest which, in our view, have not yet received adequate attention. One of the most important concerns the choice of orbitals in Ψ_0. It has been assumed that the Hartree–Fock approximation leads to orbitals which at least provide a good starting point for a discussion of electron correlation; but there does not appear to be any general agreement on this point (Szasz, 1963; Sinanoğlu, 1964) and it is of great interest to investigate other possibilities. The approach of Section V stresses the feasibility of using orbitals which are exact or approximate eigenfunctions of any "model" Hamiltonian, and that the usual Hartree–Fock starting point has no special merit provided single excitations are admitted.

The admission of triple- and higher-order excitations (Sinanoğlu, 1961, 1962, 1964), which contribute extensively in the third and higher orders of perturbation theory, also has implications which have not yet been fully assessed. Even when only pair excitations are admitted, the spin-coupling problem becomes considerably more complicated in higher order (Steiner, 1965), particularly for open-shell systems. The equations for different excited functions are no longer uncoupled, and it is not at all clear which of these functions may reasonably be neglected.

Finally, the whole question of the use of *localized* orbitals seems to require much more attention. The use of "separated" pairs (Hurley *et al.*, 1953; Parks and Parr, 1958; McWeeny and Ohno, 1960), is particularly useful in discussing the bonds in saturated molecules where the electron pairs may be assigned unambiguously to essentially nonoverlapping regions in space. Edmiston (1963) has pointed out that when pair excitations are limited to such pairs, using a localized orbital formulation, Sinanoğlu's approach essentially coincides with that of separated pair theory. It would be particularly interesting to start from a Ψ_0 having the form of a general group function (McWeeny, 1959, 1960)—a function which may be set up very easily (McWeeny and Ohno, 1960; McWeeny and Klessinger, 1965)—and to examine in more detail the separation of the correlation energy into intragroup and intergroup components.

REFERENCES

BRUECKNER, K. A. (1955). *Phys. Rev.* **100**, 36.

EDMISTON, C. (1963). *J. Chem. Phys.* **39**, 2394.

GOLDSTONE, T. (1957). *Proc. Roy. Soc.* (*London*) **A239**, 267.

HURLEY, A. C., LENNARD-JONES, J. E., and POPLE, J. A. (1953). *Proc. Roy. Soc.* (*London*) **A220**, 446.

HYLLERAAS, E. A. (1929). *Z. Physik* **54**, 347.

JAMES, H. M., and COOLIDGE, A. S. (1933). *J. Chem. Phys.* **1**, 825.

KRAUSS, M., and WEISS, A. W. (1964). *J. Chem. Phys.* **40**, 80.

LINDERBERG, J. (1960). *J. Mol. Spectr.* **5**, 1.

MCWEENY, R. (1959). *Proc. Roy. Soc.* (*London*) **A253**, 242.

MCWEENY, R. (1960). *Rev. Mod. Phys.* **32**, 335.

MCWEENY, R. (1962). *Phys. Rev.* **126**, 1028.

MCWEENY, R. (1964). "Molecular Orbitals in Chemistry, Physics, and Biology," p.305. Academic Press, New York.

MCWEENY, R., and KLESSINGER, M. (1965). *J. Chem. Phys.* **42**, 3343.

MCWEENY, R., and OHNO, K. (1960). *Proc. Roy. Soc.* (*London*) **A255**, 367.

MCWEENY, R., and SUTCLIFFE, B. T. (1963). *Proc. Roy. Soc.* (*London*) **A273**, 103.

PARKS, J. M., and PARR, R. G. (1958). *J. Chem. Phys.* **28**, 335.

RODBERG, L. S. (1957). *Ann. Phys.* (*N.Y.*) **2**, 199.

ROOTHAAN, C. C. J. (1960). *Rev. Mod. Phys.* **32**, 179.

SINANOĞLU, O. (1961). *Proc. Roy. Soc.* (*London*) **A260**, 379.

SINANOĞLU, O. (1962). *J. Chem. Phys.* **36**, 706 and 3198.

SINANOĞLU, O. (1964). *Adv. Chem. Phys.* **II**, 315.

STEINER, E. (1965). To be published.

SZASZ, L. (1959). *Z. Naturforsch.* **14a**, 1014.

SZASZ, L. (1960). *Z. Naturforsch.* **15a**, 909.

SZASZ, L. (1962a). *Phys. Rev.* **126**, 169.

SZASZ, L. (1962b). *J. Math. Phys.* **3**, 1147.

SZASZ, L. (1963). *Phys. Rev.* **132**, 936.

Quantum Chemistry and Crystal Physics

Stability of Crystals of Rare Gas Atoms and Alkali Halides in Terms of Three-Atom and Three-Ion Exchange Interactions

LAURENS JANSEN

International Division
Battelle Memorial Institute
Geneva, Switzerland

I. Background of the Stability Problems

A. Introduction

An essential problem in solid state physics and crystal chemistry concerns the explanation of observed crystal structures and, as the case may be, the interpretation of transitions between different structures exhibited by one and the same chemical compound. The interest in this subject, fascinating as it is *per se*, also originates from the expectation that a solution, for at least a few simple types of solids, would entail a profound insight in the mechanisms which govern the interactions between atoms or molecules in dense media. In this respect, the study of solids offers a decisive advantage compared to that of liquids and dense gases, where the necessity of evaluating configurational distribution functions already constitutes a problem of formidable complexity.

In the history of the stability problem two classes of solids, namely, those of the alkali halides and of the rare gas atoms, have, due to their simplicity, received by far the most attention in the literature. In the following we will deal almost exclusively with these two cases, also in view of the fact that recent further results pertain most directly to them. Both for the rare gas crystals and for the solids of the alkali halides the first explicit stability considerations were given approximately 40 years ago; for reviews of the earlier work on rare gas crystals we refer to articles by Dobbs and Jones (1957) and by Pollack (1964). A detailed exposé of early work on the stability problem for alkali halide crystals is to be found, for example, in the well-known treatises by Born and Huang (1954, Chapters 1 and 3) and by Pauling (1960, Chapter 13), as well as in a recent article by Tosi and Fumi (1962). We will review the main aspects in Sections I,B and I,C.

Let us first summarize the experimental data on rare gas and alkali halide crystals. There is good experimental evidence (Pollack, 1964) that neon, argon, krypton, and xenon crystallize in a face-centered cubic (fcc) lattice, whereas helium (He^4), under pressure, exhibits hexagonal close-packed (hcp) structure (Keesom and Taconis, 1938). At higher pressures and temperatures He^4 appears to exist in the face-centered cubic structure (Dugdale and Simon, 1953). Recently, a third solid phase of He^4 was found which is probably body-centered cubic (cf. Schuch *et al.*, 1963). Further, recent experiments on solid argon have revealed that sometimes a *metastable* solid phase exists which is hexagonal close-packed (Barrett and Meyer, 1964; Meyer *et al.*, 1964). Upon the addition of small amounts of nitrogen this phase becomes stable near the melting point.

Apparently, then, solid helium behaves quite differently from the other rare gas crystals, and the problem is simply to find out why this is the case.

Concerning solids of the alkali halides, there is conclusive evidence (Born and Huang, 1954; Pauling, 1960) that all of them except three crystallize at normal pressures and temperatures in the so-called sodium chloride configuration, consisting of two interpenetrating face-centered cubic lattices, each occupied by one kind of ion. The three exceptions are cesium chloride, cesium bromide, and cesium iodide, which exhibit the so-called cesium chloride structure; in this configuration the two interpenetrating lattices are simple cubic. In addition, all rubidium halides, as well as all potassium halides except its fluoride, undergo pressure transitions from the sodium chloride to the cesium chloride structure. Recently, Evdokimova and Vereshchagin (1963a,b) have established the presence of small amounts ($<8\%$) of the CsCl-type structure in sodium chloride at a pressure of approximately 18 kbar and under exposures up to 20 hours. However, this observation concerns only the possible *onset* of a transition. Remarkably enough, such a transition has so far not been found with cesium fluoride.

The problem here is to explain the different behavior for the three cesium halides and to attempt to account theoretically for experimentally observed transitions.

B. Rare Gas Crystals. Stability Studies Based on Two-Atom Interactions

We will now start with an analysis of the stability problem for rare gas crystals. First of all, we note that the cohesive energies of these solids are very small compared to those of ionic crystals or of metals, namely, only of the order of a few kcal kilocalories per mole, i.e., of the order of 1/10 eV per atom. The ionization potentials of rare gas atoms, on the other hand, lie between 10 and 20 eV, so that the interactions are very weak indeed. The sublimation energies of metals are ten to fifty times larger and the lattice energies of alkali halide crystals are of the order of 100 times larger than the cohesive energies of rare gas solids.

It is further known that the interaction potential between two rare gas atoms consist of a weak van der Waals attraction at large distances and a steep repulsion at small distances, of the form as given approximately in Fig. 1. It is important to realize that our knowledge concerning the pair-potential $E(R)$ is largely *empirical*; it is usually obtained by assuming a certain analytic form for $E(R)$ with adjustable parameters, and by fitting experimental PVT data at low densities, or viscosity data, combined with crystal properties. The difficulty here is that low-density gas data are necessarily rather insensitive with respect to the precise form of $E(R)$, whereas the analysis of high-density gas data is statistically much too complicated to yield accurate information. Crystal properties give only information about $E(R)$ in the neighborhood of the potential minimum and their analysis contains, in addition, the inherent assumption that the interactions may be described by means of pair-potentials only. For details, we refer to the treatise by Hirschfelder *et al.* (1954, Chapter 8) and to a recent analysis by Munn (1964).

From theory it has been known since London's famous calculation (London, 1930b) of 1930 that the long-range part of $E(R)$ can be obtained from second order of perturbation theory. It starts with an R^{-6} dependence at large distances (we do not consider retardation forces) which represents induced dipole-dipole interactions, changing to R^{-8}, R^{-10}, ..., as R becomes smaller (higher multipoles), but soon the charge clouds start to overlap appreciably and the interactions assume *exchange* character. These exchange forces predominate at the smallest distances; they then give rise to repulsion and appear already in first order of perturbation theory.

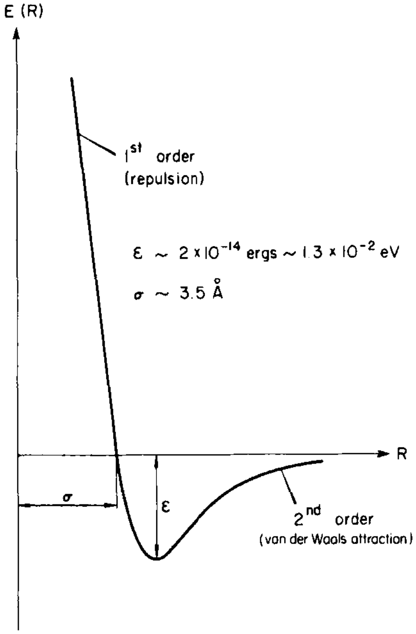


Fig. 1. Approximate pair-potential $E(R)$ between neon, argon, krypton, or xenon atoms, as a function of the interatomic distance R.

Except for the smallest atoms (helium) the accurate evaluation of these repulsive interactions is beyond the reach of present-day computational techniques.

In view of these uncertainties concerning $E(R)$ it would seem hopeless at first sight to give an explanation of the crystal structures of rare gas atoms from first principles, but fortunately the situation is more favorable. First of all, we have only to compare the two close-packed structures fcc and hcp, and the differences between these two structures permit certain general conclusions regarding their relative stability.

In general, at a temperature T and a pressure P, that structure is the most stable one which has the lowest free enthalpy (Gibbs free energy) $G = U - TS + PV$, where U denotes the crystal energy and S its entropy. If we neglect volume changes in transition, then we may compare, instead, the Helmholtz free energies $F = U - TS$. It is convenient to consider first the absolute zero of temperature, where $F(0) = U(0)$, and to split the energy $U(0)$ in two parts: the static lattice energy $U_{\text{st}}(0)$ and the zero-point energy $U_{\text{zp}}(0)$. We consider these three different parts of the free energy in the order (*a*) $U_{\text{st}}(0)$, (*b*) $U_{\text{zp}}(0)$, and (*c*) the thermal energy $U(T) - U(0) - TS$.

1. *The Static Lattice Energy* $U_{\text{st}}(0)$

We stated before that the stability analysis for rare gas crystals may be restricted to a comparison between the face-centered cubic and hexagonal close-packed lattices only. As far as pair-interactions between the atoms are concerned, the validity of this limitation was proven by Born and collaborators (Born, 1940, 1942, 1944; Misra, 1940; Power, 1942). They considered the static lattice energy only, and an interatomic potential of the form

$$E(R) = A/R^n - B/R^m; \qquad n > m, \quad A \text{ and } B > 0. \tag{1}$$

These authors showed that the simple-cubic structure is always unstable, that the body-centered lattice is very likely to be unstable, whereas the face-centered cubic configuration is always a stable one. For the hexagonal Bravais lattice compared with the hexagonal close-packed configuration they found that the first one, for a law of force of type (1), is very unlikely to be stable, whereas the hexagonal close-packed configuration satisfies the stability conditions. Similar considerations were given by Nabarro and Varley (1952) for hexagonal structures with an additional interaction energy depending only on the volume of the solid (like the Fermi energy of free electrons). Earlier, Lennard-Jones and Ingham (1925) had already compared the static lattice energies of the simple-cubic, body-centered cubic, and face-centered cubic lattices with a pair-potential of the form shown in Eq. (1). An explicit assumption in all these calculations is that the lattice energy can be written as a sum of interactions between *isolated* pairs of atoms, i.e.,

$$U_{\text{st}} = \sum_{i<j} E(R_{ij}), \tag{2}$$

where R_{ij} is the distance between atoms i and j, isolated from the rest.

Kihara and Koba (1952; Kihara, 1953) calculated the difference in static lattice energy ΔU_{st}, between the hcp and fcc structures, using for $E(R)$ a Lennard-Jones $(s, 6)$ potential

$$E(R) = \frac{\varepsilon}{s-6}\,[6(r_0/R)^s - s(r_0/R)^6], \tag{3}$$

where $\varepsilon > 0$ is the depth of the potential minimum at $R = r_0$, with variable s, and for a modified Buckingham potential ("exp-six" potential)

$$E(R) = \frac{\varepsilon}{a-6}\,[6 \exp a(1 - R/r_0) - a(r_0/R)^6], \tag{4}$$

with different values for the parameter a. For $E(R)$ of the form shown in Eq. (3) Kihara and Koba found that ΔU_{st} always favors the hexagonal lattice, for s between 7 and ∞, whereas for the exp-six potential the hexagonal structure is more stable for $a > 8.675$, and the cubic for $a < 8.675$. In all cases the relative difference $\Delta U_{st}/U_{st}$ is only of the order of 10^{-4} (one hundredth of 1%), remarkably constant with respect to a variation of the parameters s and a.

An explanation of the stability of the cubic structure on the basis of an exp-six potential with $a < 8.675$ must, however, be discarded for the following reasons:

(*i*) experimental values of a, determined from data on dilute gases, range from 12 to 14.5 for the rare gas atoms (Hirschfelder *et al.*, 1954)

(*ii*) the limiting value is too close to ~ 8.5, where the exp-six potential ceases to exhibit a minimum at finite distance between the atoms.

Similar calculations have been made by Barron and Domb (1955), with a more general Lennard-Jones (n, m) potential, i.e., of the form shown in (1), with quite similar results.

At least qualitatively, the above results can easily be understood without making any explicit calculations, as has been indicated by Professor L. Onsager (unpublished data, see Barron and Domb, 1955). Both the fcc and hcp structures can be built by stacking two-dimensional hexagonal arrays of atoms on each other. If we call a first layer A and a second layer B, then for the third layer there are two possibilities: *either* another layer A *or* a layer C rotated through 60° with respect to A.

In the first case we obtain the sequence ABABAB ..., which represents the hexagonal close-packed lattice. In the second case the sequence is ABCABC ..., representing the face-centered cubic configuration. Whereas there is no difference in the interactions between AB, AC, or BC layers, two A layers attract each other slightly more than a layer A and a layer C, for sufficiently short-ranged attractions, since the distance between corresponding atoms in two A layers is a minimum. In trying to solve the stability problem for rare gas crystals we must understand why, *in spite of the (slight) loss in static pair energy of the solid, the sequence ABC is energetically more favorable than the sequence ABA.*

2. *The Zero-Point Energy* $U_{zp}(0)$

We next consider the effect of zero-point energy on the stability of the two close-packed structures. However, zero-point energy cannot reasonably be expected to sensitively affect the relative stability of the fcc and hcp structures since solid He^4, which has the largest zero-point energy, crystallizes in the hcp configuration. For this reason, Jansen and Dawson (1954, 1955) used the approximation proposed by Corner (1939). In the model Corner uses, the atoms carry out harmonic oscillations about their equilibrium positions. The frequency of oscillation is calculated for one atom with the other atoms at rest. The following expression for the zero-point energy is then obtained for a Lennard-Jones $(s, 6)$ potential:

$$U_{zp}/\varepsilon = F(s)\Lambda^*[s(s-1)\, r_0/R_0)^{s+2}C_{s+2} - 5s(r_0/R_0)^8 C_8]^{1/2}, \qquad (5)$$

with

$$F(s) = (9/8)(5/3)^{1/2}/\{[2\pi^2(s-6)]^{1/2}(s/6)^{1/(s-6)}\}$$

and

$$\Lambda^* = h/\sigma(m\varepsilon)^{1/2}.$$

The distance between two atoms for zero potential is σ, whereas ε is the depth of the potential well, at distance r_0. Further, R_0 is the nearest-neighbor distance in the crystal, m is the mass of the atom, and the C_s are crystal sums for the fcc lattice, tabulated by Lennard-Jones and Ingham (1925) and by Kihara and Koba (1952; Kihara, 1953). In case of the hcp lattice, analogous equations hold, with C_s replaced by the hexagonal crystal sums H_s. Values for H_s were determined by Goeppert-Mayer and Kane (1940) and Kihara and Koba (1952; Kihara, 1953).

The range $\Lambda^* = 0.1$ to 0.7 covers all rare gases except helium. The zero-point energy of helium is so large that it cannot be calculated on the basis of Corner's model of harmonic oscillations; we will therefore not discuss solid helium in this connection. For the exp-six potential the expression for the zero-point energy is:

$$U_{zp}/\varepsilon = G(a)\Lambda^*\left(\frac{a}{a-6}\right)\left[\left(\frac{a}{2} - \frac{1}{R^*}\right)e^{a(1-R^*)} - 5(C_8/R^{*8})\right]^{1/2}, \qquad (6)$$

where

$$G(a) = (9/8)(5/3)^{1/2}/\{[2\pi^2 a/(a-6)]^{1/2}x(a)\}$$

and where $x(a) = r_0/\sigma$ is a solution of the transcendental equation

$$6 \exp(a - a/x) = ax^6;$$

R^* is the reduced lattice distance R_0/σ. Values for $G(a)$ can be computed from a table given by Hirschfelder and Rice (1954) for the range $a = 12.0$ to 15.0. For the hcp lattice the expression for the zero-point energy is the same as (6), except that C_8 is replaced by H_8.

The expressions (5) and (6) for the zero-point energy are then added to those for the static lattice energy calculated on the basis of the potentials (3) and (4), respectively. Then these sums are differentiated with respect to R_0, whereupon U_{st} and U_{zp} are evaluated for minimum energy. It then appears (Jansen and Dawson, 1954, 1955) that, for all admissible values of the parameters s and a, the hexagonal lattice is again the more stable one. The relative difference is again of the order of 0.01%, i.e., of the same order as that obtained when we neglected zero-point energy altogether.

Barron and Domb (1955) carried out a more accurate calculation of the zero-point energy for the two close-packed structures, using the Born–von Karman lattice dynamics theory. The application of lattice dynamics to a comparison between different crystal structures is straightforward, as discussed in detail by Begbie and Born (1947). For nearest-neighbor interactions only Barron and Domb find that the fcc lattice is favored by about 0.01% of the zero-point energy. Upon extending the calculations to interactions between all neighbors and using an approximate expression for the zero-point energy given by Domb and Salter (1952), the relative difference in zero-point energy becomes of an order not greater than 0.01%, but its sign is uncertain. A certain conclusion is, however, that zero-point energy has no decisive effect on the relative stability of the two close-packed configurations.

Recently, Wallace (1964) carried out a lattice dynamics calculation of the crystal energy at absolute zero, whereby he took into account also *anharmonic* contributions, on the basis of a Lennard-Jones (n, m) potential. For the harmonic part of the zero-point energy, the results agree well with those of Barron and Domb. As far as anharmonicity is concerned, Wallace concludes that such effects would be sufficient to overcome the 0.01% barrier in U_{st} for neon and argon, but not for krypton and xenon. It is clear, therefore, that the stability of the fcc configuration for heavy rare gas crystals is not to be sought in a difference between zero-point energies for the two structures.

We conclude that, on the assumption of pair-interactions, all rare gases should crystallize in the hexagonal close-packed configuration at absolute zero. Of course, all experimental data refer to finite temperatures, so that it *could* in principle happen that the crystals are actually hcp at absolute zero, as predicted, but that a transition to the fcc structure occurs at a temperature below the lowest experimental value. Although this possibility is unlikely to be realized since the heavy rare gas crystals are known to be fcc as low as 4.2°K, we should examine it in some detail.

3. *The Thermal Energy* $U(T) - U(0) - TS$

To investigate the possibility of a thermal transition between the hexagonal and cubic lattices, we must calculate the variation with temperature of their difference in Helmholtz free energy, starting from $T = 0$. In the limit of very low temperatures the specific heat at constant volume C_V of the lattice varies according to the Debye T^3 law

$$C_V = b(T/\Theta)^3,$$

where Θ is the (Debye) characteristic temperature of the crystal (in the limit $T = 0$), and where b is a material constant. We now obtain for the Helmholtz free energy of either structure:

$$\begin{aligned} F(T) = U(T) - TS(T) &= U(0) + \int_0^T C_V\,dT' - T\int_0^T \frac{C_V}{T'}\,dT' \\ &= U(0) - \frac{b}{12}\,T(T/\Theta)^3; \end{aligned} \tag{7}$$

the second member on the right represents the *thermal energy* of the crystal. The characteristic temperature Θ, depending on the maximum frequency of acoustic waves, ν_m, according to $\Theta = h\nu_m/k$, will be different for the two lattices. Denoting quantities for the hexagonal lattice by a subscript h, and those for the cubic lattice by a subscript c, we obtain

$$\Delta F(T) \equiv F_h(T) - F_c(T) = \{U_h(0) - U_c(0)\} - \frac{b}{12}\,T^4\{1/\Theta_h^3 - 1/\Theta_c^3\}. \tag{8}$$

At the transition temperature, the left-hand side of this equation is zero. From the two-body calculations we know that $\{U_h(0) - U_c(0)\}$ is negative. Thus, in order for a transition from the hexagonal to the cubic lattice to occur as the temperature is raised, it is necessary that

$$\Theta_h > \Theta_c \qquad \text{(condition for thermal transition).}$$

Barron and Domb (1955) have confirmed the validity of this inequality from their lattice dynamics analysis. The calculated transition temperatures, however, appear to be comparable with the *melting points* of rare gas crystals; i.e., they are much too high to provide an explanation for the stability of the fcc configuration.

The results of Sections 1, 2, and 3 on the static lattice energy, zero-point energy, and the possibility of thermal transitions between the two close-packed structures lead us to the following conclusion.

The stability of the face-centered cubic configuration for heavy rare gas crystals cannot be explained on the basis of pair-interactions which are of the central force type.

There are, apparently, only two possibilities left.

(*a*) The interactions between rare gas atoms in dense media are essentially of two-body type, but they deviate from central forces. This implies that the charge distributions in the isolated atoms are *not* spherically symmetric.

(*b*) Interactions between *more than two atoms* are decisive for the determination of the stable crystal structure. We consider, in the first place, *three-atom* interactions; the differences between two crystal structures are then given not only in terms of differences in the arrangements of pairs, but also in terms of differences in triplet configurations.

A third possibility, namely, that the observed face-centered cubic structures are actually *metastable* states of rare gas crystals, can now be discarded on the basis of the experiments with metastable *hexagonal* argon (Barrett and Meyer, 1964; Meyer *et al.*, 1964), which we mentioned before.

a. Noncentral, Two-Body Forces. Cuthbert and Linnett (1958) have introduced noncentral forces between two rare gas atoms on the basis oj the following consideration. The helium atom contains only *s* electrons, whereas the outer shells of the other rare gas atoms contain electrons in the *ns* and the three *np* orbitals, with $n = 2, 3, 4$, and 5. Linnett and Poë (1951) had reasoned that, because of spin correlation, the most probable configuration of the group of eight electrons in heavy rare gas atoms involves s^2p^6 hybridization with four pairs of electrons at the corners of a regular tetrahedron, the nucleus being at the center. Electrostatic interactions involving high electric (permanent) multipoles are then introduced which, according to Cuthbert and Linnett, favor the fcc configuration for heavy rare gas crystals.

To this suggestion we remark that:

(*i*) no estimate concerning the order of magnitude of this correlation effect has been made;

(*ii*) the interactions between nearest neighbors in the crystals have considerable *exchange* character;

(*iii*) the four electrons with positive spin are independent of the four electrons with negative spin (as far as spin-correlation is concerned). Hence, Coulomb correlation would tend to favor *two* tetrahedra, the corners of one occupied by positive-spin electrons, the corners of the other by negative-spin electrons (Pauling, 1960, Chapter 4, pp.128–130). These two tetrahedra are inverted with respect to each other.

It is of interest to note that the idea of *pre*polarized spins in an isolated atom of the heavy rare gas atoms has recently been considered by Allen (1962, 1963) in a discussion of possible binding mechanisms in xenon-fluorine compounds. However, in the case of *very weak* interactions which we are considering in rare gas crystals such correlations constitute only a fine structure which is of no importance for crystal stability. For comparison, we mention that the energy of the xenon-fluorine bond corresponds to seventy times the depth of the potential well between two xenon atoms. The occurrence of polarized spins under *strong* perturbations is known from the MO formulation of ligand field theory (Ballhausen, 1962, Chapter 7) and from the description of delocalized molecular orbitals to account for the geometry and nuclear quadrupole constants of the polyhalide ions (cf. Jortner *et al.*, 1963). The analogy between xenon-fluorine compounds and polyhalide ions seems to be of particular interest (Rundle, 1963; Pitzer, 1963) in this respect.

Two different mechanisms which have been proposed to explain the stability of the cubic structure also invoke electrostatic interactions between electric multipoles, although they are strictly speaking many-atom effects. Kihara (1960) considered interactions between *induced* electric multipole moments of the atoms; these moments are induced by the crystal field. For reasons of symmetry, the first nonvanishing induced moment for an hcp lattice is an eight-pole (octupole), and a sixteen-pole (hexadecapole) in case the lattice is cubic. The electrostatic interactions between such induced multipole moments are always repulsive, decreasing with increasing multipole order. For this reason, the fcc lattice is qualitatively favored. Very recently Knox and Reilly (1964) have made a detailed analysis of Kihara's suggested explanation. They find that, from a purely electrostatic point of view, Kihara's multipole-multipole interactions are in fact *not present* in the expression for the cohesive energy of the crystal. In addition, a much larger interaction (monopole-multipole) exists which Kihara did not consider. This interaction is still quite small, even though it would favor the cubic lattice.

Prins *et al.* (1952) forwarded a suggestion for stability which is also based on general symmetry properties of the surroundings of an atom in the two close-packed structures, namely that a compression of the eight-electron shell in a cubic arrangement of atoms is more favorable than in a crystal of hexagonal symmetry. Also, according to the same authors, the *dynamics* of crystal growth might be of influence in that a cubic crystallite grows faster than a hexagonal one. However, the phenomena observed with metastable hexagonal argon appear to show that such is not the case. Finally, we mention a suggestion by Mulliken (1951) concerning the possible occurrence of polarization forces of a charge-transfer type and their effect on crystal stability. However, in view of the relatively high ionization potentials for rare gas atoms such forces are unlikely to be considerable. In addition, we require them to be sufficiently *different* for the two structures. On the basis of any type of *pair*-interactions this necessitates that their contribution for either structure must be large.

b. Many-Atom Interactions. All other explanations having failed, we now turn to the possible occurrence of many-atom interactions in rare gas crystals. This implies that Eq. (2) for the static lattice energy should be modified as follows

$$U_{\mathrm{st}} = \sum_{i<j} E(R_{ij}) + \sum_{i<j<k} E(R_{ij}, R_{ik}, R_{jk}) + \cdots, \tag{9}$$

where the second member on the right represents three-atom interactions, in principle to be supplemented by simultaneous interactions between four, five, etc., atoms.

A necessary condition which must be imposed on this kind of cluster expansion is that it converge rapidly with increasing cluster size. For this reason, three-atom interactions should give the main contribution to the many-body component of the crystal energy. This limitation agrees also with vast experience on molecular crystals, namely, that their properties can in good approximation be described by means of atomic properties and (effective) pair-potentials. For this reason, we will consider simultaneous interactions between *three* atoms only.

In order that such three-atom interactions may provide an explanation for crystal stability, two *a priori* conditions must be satisfied. First, the consideration of triplets of atoms must increase the differences between the hcp and fcc structures. Second, these differences must be large enough to overcome the 0.01% barrier in the pair-potential between the two structures *by at least an order of magnitude*. Otherwise, the stability problem cannot be regarded as solved.

The first of these two conditions is evidently satisfied. To illustrate this, we consider a central atom in the hcp and fcc structures and its twelve nearest neighbors. As far as pair-interactions are concerned, the immediate neighborhood of the atom in the two structures is, therefore, identical. Let us now consider all possible triplets of atoms formed by the central atom and any two of its twelve nearest neighbors. There are sixty-six such triplets in either structure; upon inspection it appears that fifty-seven triplets in the hcp lattice are the same as fifty-seven in the fcc configuration, *but nine are different*. From this we see that the two structures are already different for *nearest* neighbors if we consider triplet configurations.

The second condition requires that three-atom interactions must be highly *structure-sensitive*. This implies that they must be of short range, i.e., of exchange type, which is also required by the limitation of the cluster expansion, Eq. (9), to triplets of atoms.

Before going into the analysis of possible three-atom interactions, we will first review the stability of alkali halide crystals and then combine the two stability problems.

C. Alkali Halide Crystals. Stability Studies Based on Two-Ion Potentials

The development in the literature of stability studies concerning crystals of the alkali halides has been strikingly similar to that of the rare gas solids. An excellent summary of the earlier work on alkali halide crystals has recently been given by Tosi and Fumi (1962), to which paper we have already referred earlier.

As for rare gas solids we start with a consideration of the *static* lattice energy at the absolute zero of temperature. Let us compare the NaCl configuration (two interpenetrating face-centered cubic lattices) and the CsCl structure (two interpenetrating simple-cubic lattices); at normal pressures and temperatures the NaCl structure is experimentally found to be the more stable one for all alkali halides except for CsCl, CsBr, and CsI, which exhibit the CsCl configuration. Further, we note that the alkali halide ions are isoelectronic with rare gas atoms: Li^+, H^- with He; Na^+, F^- with Ne; K^+, Cl^- with Ar; Rb^+, Br^- with Kr; and Cs^+, I^- with Xe. Therefore, the interactions between the ions on one side and the atoms on the other side must be *of the same form* (interactions between closed shells of electrons) if we subtract the electrostatic interactions between the ionic charges and if we disregard polarization effects in view of the high symmetry of unstrained crystals.

The sum of interactions between all point ionic charges of the crystal is called the Madelung energy. The remaining part of the lattice energy appears to be positive (repulsion) for all alkali halides, and of the order of 10% of the crystal energy, with opposite sign. The reason why we have a repulsion between the closed shells of electrons in this case, instead of an attraction as with rare gas crystals, is because the Madelung energy compresses the solid to such an extent that neighboring ions repel each other. The nearest-neighbor distances in alkali halides are, consequently, considerably smaller than those in corresponding rare gas solids.

In first approximation we then write the potential between two nearest neighbors in the crystal (each positive ion is surrounded by a first shell of negative ions, and inversely), at distance R_0, as

$$E(R_0) = -e^2/R_0 + b/R_0^n, \tag{10}$$

i.e., in terms of the electrostatic attraction between two point charges of magnitude $\pm e$, and a repulsion between closed shells, varying as the inverse nth power with distance. This equation is analogous to the Lennard-Jones potential, Eq. (3), from which we have now dropped the van der Waals attraction.

In considering also next-nearest neighbors, which in the NaCl and the CsCl configurations are of the *same* type as the central ion, the first term on the right of Eq. (10) changes sign, the repulsion is much smaller unless the ions are very large, and we should also add a van der Waals attraction, because these like ions normally attract each other. This procedure would lead to a very complicated expression for the static lattice energy, containing several parameters. In addition, it would not be correct to compute the Madelung energy as if the ions were *point* charges, since nearest neighbors usually overlap appreciably.

Instead we assume, *empirically*, that the crystal energy may be calculated *as if* electrostatically the ions were point charges and *as if* there were no van der Waals attractions between the closed shells. Further, we contract the total repulsion in the form of a contribution between *nearest* neighbors only. If we assume *pair*-interactions only, then the expression for the energy of the lattice of N anions and N cations is

$$U_{st} = -N\alpha e^2/R_0 + NSb/R_0^n, \tag{11}$$

where α is called the Madelung constant and where S denotes the coordination number of the lattice (i.e., the number of nearest neighbors); $S = 6$ for the NaCl configuration and 8 for the CsCl structure, whereas α has the values 1.7476 and 1.7627, respectively, for these lattices.

Since the cohesive energies for the ionic crystals are of the order of one hundred times those of rare gas crystals, the zero-point energy is of comparatively little importance. In that case we may consider U_{st} itself to have a minimum at nearest-neighbor distance. Taking derivatives with respect to this distance, Eq. (11) leads to the relation

$$U_{st} = \frac{-N\alpha e^2}{R_0}\left(1 - \frac{1}{n}\right), \tag{12}$$

so that for $n = 10$ the ratio between absolute values of Madelung energy and repulsion is equal to 10, as required approximately by experiments.

Instead of a repulsion proportional to R^{-n}, Born and Mayer (1932) introduced an exponential law (Born–Mayer potential)

$$E(R_0) = -e^2/R_0 + Be^{-R_0/\rho}, \tag{13}$$

which is the analog of the "exp-six" potential between rare gas atoms. For such a potential the condition for minimum static lattice energy, Eq. (12), reads

$$U_{st} = \frac{-N\alpha e^2}{R_0}\{1 - (\rho/R_0)\}, \tag{12a}$$

from which we conclude that $R_0/\rho \approx 10$ will yield approximate agreement with experimental data.

We can now pursue further the parallel with rare gas crystals and inquire which one of the two structures (NaCl or CsCl) is more stable at the absolute zero of temperature. Hund (1925) has carried out this analysis for the potential (10), whereby he included for comparison also the zinc blende structure (two interpenetrating fcc lattices, as in the NaCl structure, but in this case the coordination number of an ion in the composite lattice is 4), for which the Madelung constant, α, is equal to 1.6381.*

In view of the difference in coordination number for the three lattice types we cannot compare the static lattice energies U_{st}, i.e., Eq. (12) or (12a), for the same values of the nearest-neighbor distance R_0. Otherwise it is the Madelung energy which determines stability completely, and we would have, in order of decreasing preference: CsCl, NaCl, ZnS. Instead, we must, from the condition for minimum U_{st}, solve for R_0 in the three

---

* ZnS occurs in two modifications: sphalerite, where the two sublattices are fcc, and wurtzite, in which these two lattices are hcp. Their Madelung constants differ by less than 0.2%, and in the Born–Mayer model of ionic solids no distinction can be made between them. Both modifications have coordination number 4.

cases and use the three different values for R_0 in the expression for U_{st}. In case of the potential (10) this gives:

$$U_{\mathrm{st}} = -\left(\frac{\alpha^n}{S}\right)^{1/(n-1)} \left\{ N\left(\frac{n-1}{n}\right)\left(\frac{e^{2n}}{nb}\right)^{1/(n-1)} \right\}. \tag{14}$$

Since the second factor on the right is the same for all structures, the relative stability depends only on the value of α^n/S; the lattice type with the highest value of this quantity is the most stable one. For very small values of n the coordination number S determines stability; the lattice with the lowest S is then the most stable one (ZnS). A more detailed analysis gives the following results:

$n < 6.3$:	ZnS,	NaCl,	CsCl,
$6.3 < n < 9.5$:	NaCl,	ZnS,	CsCl,
$9.5 < n < 33$:	NaCl,	CsCl,	ZnS,
$33 < n$:	CsCl,	NaCl,	ZnS.

We see at once that for values of n in the neighborhood of 10 the NaCl structure is the most stable one and that only for very high values of n is the CsCl structure preferred. Since we know from experiments that n must have a value of approximately 10 in the present model, the Born–Mayer potential in this simple form *fails to account for the observed stability of the cesium chloride configuration for CsCl, CsBr, and CsI.*

Attempts to correct the model have taken two directions. (*a*) The first type is based on vast experimental evidence that the *ratio* between the "size" of cation and anion plays an important role for stability. Empirical rules for predicting the stable crystal structure of ionic and other solids from a knowledge of ionic or atomic radii alone were formulated by Goldschmidt (see Born and Huang, 1954). The effect of ionic size can be taken into account most explicitly on the basis of a model of *rigid spheres* with point-charge interactions only; the parameters are then α, S, and the ratio (r_+/r_-) between the sizes of cation and anion. We then obtain the following stability rules:

$0.717 < r_+/r_-$:	CsCl,	NaCl,	ZnS,
$0.609 < r_+/r_- < 0.717$:	NaCl,	CsCl,	ZnS,
$0.325 < r_+/r_- < 0.609$:	NaCl,	ZnS,	CsCl,
$r_+/r_- < 0.325$:	ZnS,	NaCl,	CsCl.

Although also this model of rigid spheres is deficient in several respects, it strikingly confirms experimental evidence that the ZnS structure only occurs for small values of the ratio between cation and anion for the silver halides (AgI), but this rule appears not to hold for the alkali halides (e.g., NaI has NaCl structure). In any case, such model calculations show that the ratio of ionic sizes is indeed an important parameter in determining the stable crystal structure.

(*b*) The second type of suggested corrections to the original Born–Mayer potential involved the addition of further terms to Eq. (10) or Eq. (13), with negative sign. These terms are supposed to correct for the omission of *van der Waals attractions* between the closed shells of electrons on the ions. Such a modification appeared to be called for in order to diminish the energy difference between the NaCl and the CsCl configurations, since attractive forces favor the CsCl structure somewhat.

However, inclusion of van der Waals attractions at large distances in an empirical potential function does not necessarily render this potential more "realistic" at small distances. In addition, the *net* effect on the crystal energy amounts only to a fraction of the sum of van der Waals interactions for the ions (Born and Huang, 1954; Chapter 1, p.28). This is because the crystal is compressed somewhat by the attractions and the repulsive forces increase considerably and counteract the contribution of the attractive forces. It appeared that, in order to obtain a significant *net* difference between the two structures, very large van der Waals attractions would be necessary. Born and Mayer (1932) used the free-ion values for the van der Waals forces, but under those conditions CsCl was found to be more stable in the NaCl structure. Later, May (1937, 1938) repeated the same calculations, using larger van der Waals coefficients proposed by Mayer (1933), but the result was still negative. Huggins and Mayer (1933) suggested different values for the repulsive parameters; Jacobs (1938) used these parameters plus the Mayer values of van der Waals interactions in calculating the transition pressures for the potassium and rubidium halides, but the discrepancies with experimental values persisted.

Finally, contributions to the crystal energy arising from interactions between multipole moments of the ions, induced by the crystal field, have been considered in detail by de Wette (1959). He finds that this polarization energy involves multipoles of high order, that the leading term varies with the inverse tenth power of the distance, and that the total contribution is only of the order of 1 % of the Madelung energy. Therefore, such polarization effects are much too small to significantly influence stability. We note here a direct parallel with Kihara's multipole interactions in rare gas crystals (Kihara, 1960; Knox and Reilly, 1964), although in the latter case the polarization energy is even very much smaller than for ionic crystals.

In view of the above results, Tosi and Fumi (1962) recently concluded that the only possible phenomenological approach to the stability of alkali halide solids must involve the assumption that somehow the parameters B and ρ of the Born–Mayer potential are *structure-dependent*, i.e., that they have different values for the NaCl and CsCl configurations. With this modification they showed that consistent agreement can be obtained with the observed values for the work involved in the pressure transitions of potassium and rubidium halides and for the heat absorbed in the observed thermal transition of cesium chloride at 445°C from the CsCl to the NaCl structure. Best agreement is obtained if the ρ values for the two phases differ by between 4 and 9% for the potassium and rubidium chlorides, bromides, and iodides.

The question then arises regarding the source of this dependence of ρ and B upon the crystal structure. Now it is clear that *in principle* the repulsive part of the Born–Mayer potential must be structure-dependent since it represents a *weighted* average of interactions between a central ion and its first few shells of neighbors. This average depends, therefore, also on the arrangement of ions beyond the first shell, which arrangement is different for different structures. However, this cannot possibly be an important effect, since repulsive forces decrease very rapidly with distance, and the energy differences between the two structures are large, namely, of the order of a few kilocalories per mole. Tosi and Fumi have verified this explicitly by writing the repulsion in the form

$$E_{\text{rep}}(R) = SB\exp(-R/\rho) + S'B'\exp(-R'/\rho') + \cdots, \tag{15}$$

where S and S' are the numbers of first and second neighbors, respectively, and where B, ρ and B', ρ' are the repulsive parameters associated with the first and second shells; R and R' are the distances between a central ion and the first and second shells of neighbors, respectively. If we assume that B, B' and ρ, ρ' are the same for the NaCl and CsCl structures, then fitting the differences in lattice energy between these two phases at the transition pressure and at 298°K for the potassium and rubidium halides leads to meaningless values of these parameters (B' assumes negative values, whereas $\rho' \sim 1.8\rho$). This source of structure dependence has thus clearly to be discarded.

Consequently, we now *fall back* on the same two remaining possible explanations for crystal stability as discussed in the previous section for the rare gas crystals, namely (*a*) *two-body, noncentral forces* and (*b*) *many-ion interactions*. In view of the large energy differences involved in this case, possible noncentral interactions between closed shells of electrons cannot play any role for ionic crystals. This leads us, finally, to the conclusion that *simultaneous interactions between more than two ions are likely to play an* *important role in determining the stable structure for crystals of the alkali halides.*

As with rare gas crystals, we require that the cluster expansion (9) for the static lattice energy converge rapidly. For this reason, three-ion interactions should give the main contribution to the many-body component of the crystal energy. In the following, only interactions between triplets of ions will be considered. It should be noted that the differences in arrangements of three ions, of which two are nearest neighbors of the central ion, are now much larger than those for rare gas atoms in their solids. The number of nearest neighbors in the NaCl structure is six, in the CsCl structure eight, so that we have fifteen and twenty-eight of such triplets in the two respective configurations. That these differences must be much larger is necessary because of the relatively high barrier between the two crystal structures. Again, three-ion interactions must be highly structure-sensitive; i.e., they must be of short range (exchange type), etc.

For simplicity we now assume that the three-atom interactions in rare gas crystals and the three-ion interactions in alkali halide crystals have the *same physical origin.* This assumption is based on the fact that these forces must have similar properties and, further, on the observation that ions of the alkali halides are *isoelectronic* with the rare gas atoms, both consisting of closed electron shells. Consequently, their interactions must be of the same form, if we subtract purely electrostatic forces between the ion charges from the outset and disregard polarization effects in view of the high symmetry of unstrained ionic crystals. For ionic crystals this assumption implies that, except through the Madelung energy, *the ionic charges do not play an essential role in determining the stable structure.*

We can now combine the two stability problems and start with a general analysis of possible many-atom or many-ion interactions in solids.

II. Possible Many-Atom and Many-Ion Interactions in Crystals

A. Many-Body Interactions between Rare Gas Atoms

Since the interactions between rare gas atoms, at the distances between neighbors in their solid states, are extremely weak compared to their ionization potentials, it is most convenient to base a discussion of many-atom interactions on a perturbation treatment, with free-atom wave functions as zeroth order. The first consideration of many-body interactions between rare gas atoms was given by London (1930a), who pointed out that these interactions are strictly additive in pairs, up to and including second order of perturbation theory, if the atoms are sufficiently far apart, so that the zeroth-order wave function can be written as a *simple product* of atomic wave functions.

For the first order of perturbation theory this conclusion is valid, since the perturbation Hamiltonian, H', for a system of atoms consists of pair-interactions between point charges, and the first-order energy is just the expectation value of H'. For the second-order energy this conclusion is not so obvious. Here we have to evaluate expressions of the form

$$(H')_{0\kappa}(H')_{\kappa 0}, \tag{16}$$

where κ is an (any) excited state of the system and "0" is its ground state. We assume that the zeroth-order wave function, with sufficient accuracy, is represented by a simple product of wave functions for the isolated atoms.

Consider, in the matrix element $(H')_{0\kappa}$, a term r_{12}^{-1} of H', where 1 and 2 are electrons on different atoms, a and b, say. Since these atoms are far apart, we can expand r_{12}^{-1} as a double series of electric multipole moments (operators) of atom a and atom b. If now, in the mirror matrix element $(H')_{\kappa 0}$, we consider a term r_{13}^{-1}, where electron 3 is on a *third* atom c, then κ is an excited state only for atom a. The product $(H')_{0\kappa}(H')_{\kappa 0}$ can then be written as a series of which each term contains a product of expectation values for electric multipole moments of atoms b and c of varying order. Since, however, the charge distributions of the atoms are spherically symmetric, these expectation values *vanish identically* in all orders (Hirschfelder *et al.*, 1954, Chapter 8; Jansen, 1957, 1958). Therefore, if we break the series off, as is always done, after a finite number of terms, the conclusion is that a third atom cannot interfere in expressions of type (16). However, we have, strictly speaking, in each term of (16) an *infinite* series and such an infinite sum is not necessarily zero, even if each term vanishes. This is reflected in the fact that the expectation value of r_{12}^{-1}, for example, is not precisely zero, owing to the circumstance that the charge distribution of each atom extends to infinity. A calculation of such electrostatic three-atom interactions in second order of perturbation theory with a simple-product type of wave function has been carried out by Wojtala (1964). He finds nonnegligible three-atom interactions for interatomic distances comparable with the size of the interacting atoms.

On the other hand, the results are rather formal, since a simple-product type of wave function is never strictly valid in view of the antisymmetry requirement. Since we are here interested in the effect of many-atom or many-ion interactions in dense media (solids), it is a general rule that *truncated multipole expansions and simple-product type wave functions should be avoided for their evaluation.*

For the moment we will stay with these approximations, for historical reasons. If we consider *third* order of perturbation theory, then the expressions to be evaluated are of the form

$$(H')_{0\kappa}(H')_{\kappa\lambda}(H')_{\lambda 0}, \tag{17}$$

where κ and λ are two (any) excited states of the system. Let us consider the terms r_{12}^{-1}, r_{13}^{-1}, and r_{23}^{-1} in the three perturbations H', respectively. We see that each electron index occurs now *twice* and this means that such three-atom terms will *not* vanish, whether or not we use a multipole expansion and whether or not the zeroth-order wave function is antisymmetric.

For a simple-product type of wave function and a multipole expansion of H' the first nonvanishing term for a general triplet (abc) of atoms represents the three-body component of the interaction between three induced dipole moments. This term, which is called the *triple-dipole effect*, was evaluated by Axilrod and Teller (Axilrod, 1949, 1951a; Axilrod and Teller, 1943); it constitutes a straightforward extension of the London–van der Waals R^{-6} dipole-dipole interactions between two atoms at large distances. If we denote the angles of the triangle formed by the triplet (abc) by γ_1, γ_2, and γ_3 and the distances between the atoms by R_{ab}, R_{ac}, and R_{bc}, then the triple-dipole interaction, ΔE_3, for three identical atoms is

$$\Delta E_3 = (9/16)E_{\text{av}}\alpha^3(3 \cos \gamma_1 \cos \gamma_2 \cos \gamma_3 + 1)/R_{ab}^3 R_{ac}^3 R_{bc}^3, \tag{18}$$

where α is the atomic polarizability and E_{av} an "average excitation energy" per atom; E_{av} is usually equated to the first ionization potential.

On the other hand, the dipole-dipole interaction, in London's approximation, between atoms a and b is

$$(E_2^{(0)})_{ab} = -\tfrac{3}{4}E_{\text{av}}\alpha^2/R_{ab}^6. \tag{19}$$

(In principle, the two quantities E_{av} in Eqs. (18) and (19) are not necessarily the same, but we need not consider their difference here.) In view of the approximations involved in the evaluation of three- and two-atom interactions, it is more accurate to consider the three-atom interactions *relative* to the sum of pair-interactions between the atoms of the triplet. Let us denote this sum by $E_2^{(0)}$, obtained by summing Eq. (19) for the three pairs; then we have

$$\Delta E_3/E_2^{(0)} = -(3\alpha/4)[3 \cos \gamma_1 \cos \gamma_2 \cos \gamma_3 + 1] \frac{R_{ab}^{-3}R_{ac}^{-3}R_{bc}^{-3}}{R_{ab}^{-6} + R_{ac}^{-6} + R_{bc}^{-6}}. \tag{20}$$

To illustrate the properties of the triple-dipole effect, it is convenient to take as an example triplet configurations of atoms which form *isosceles* triangles. Let $R_{ab} = R_{ac} = R$ and let γ denote the opening of the isosceles triangle at atom a. The factor of Eq. (20) in square brackets then becomes $(3/2)\cos\gamma\,(1 - \cos\gamma) + 1$, which is zero for $\gamma = 117°$. Further, for an equilateral triangle ($\gamma = 60°$), $\Delta E_3/E_2^{(0)} = -(11/32)(\alpha/R^3)$, whereas for a linear symmetric array of atoms ($\gamma = 180°$) we have $\Delta E_3/E_2^{(0)} = (12/129)(\alpha/R^3)$. Substitution of known values for the nearest-neighbor distance R in solid neon, argon, krypton, and xenon, and for the atomic polarizability α, reveals that α/R^3 varies between approximately 0.005 and 0.017 from neon to xenon. Since $\Delta E_3/E_2^{(0)}$ is only a fraction of this quantity, it is clear that the triple-dipole effect constitutes only a very small correction to the van der Waals dipole interactions, even at solid state densities. In addition, we remember that application of the triple-dipole effect is not justified in dense media, because of neglect of exchange and the use of a multipole expansion. It is, however, of interest to bear the following properties of $\Delta E_3/E_2^{(0)}$ in mind:

(*i*) $\Delta E_3/E_2^{(0)}$ is *negative* for isosceles triangles with *small* opening γ;
(*ii*) it is *positive* for isosceles triangles with *large* opening γ;
(*iii*) its magnitude at $\gamma = 60°$ is larger than at $\gamma = 180°$.

Since $E_2^{(0)} < 0$, the results imply qualitatively that for triangles with small opening the triple-dipole effect constitutes a *weakening* of the attractive forces compared with an additive sum over pairs; for large opening the three-atom interactions *increase* the attraction.

Similar third-order calculations, this time based on the Drude model of harmonic oscillators for the atoms, were carried out by Muto (1943). It should be noted that the long-range third-order interactions are per se of low sensitivity with regard to the crystal structure. Axilrod (1951a,b) summed the triple-dipole effect over the hcp and fcc configurations and found that it does favor the cubic configuration, but that the difference is *too small* to overcome even the very low barrier (one hundredth of 1% of the lattice energy) in favor of the hcp structure because of pair-interactions. The total three-body interactions are in principle *not* negligible in either lattice; they would amount to 2 to 9% of the crystal energy from neon to xenon, *if* the triple-dipole effect could be applied also to near neighbors. This shows very clearly the lack of structure sensitivity of the triple-dipole effect, as mentioned above. No qualitative improvement can be expected by choosing somewhat larger numerical coefficients in the expression (18) for ΔE_3 proposed by Kihara (1958, 1963a) for the application to the stability of rare gas crystals (Danon, 1965). A variational treatment of the triple-dipole effect has been given by Midzuno and Kihara (1956). Finally, although the triple-dipole effect cannot be used at densities of solids or liquids, it should be applicable to low-density phenomena in gases. This has been confirmed for third virial coefficients by Kihara (1958, 1963a), Graber and Present (1962), and Sherwood and Prausnitz (1964).

A variety of different many-atom types of interactions, reported in the literature, can be classified in the same category as the triple-dipole effect, namely those for which the overlap of charge clouds is either neglected or treated classically (Kihara, 1960; Prins *et al.*, 1952; Ayres and Tredgold, 1955, 1956) or those based on the Drude model of coupled harmonic oscillators with dipole interactions (Bade, 1957; Bade and Kirkwood, 1957; Sparnaay, 1959; Doniach, 1963). Bade (1958) calculated, on the basis of the harmonic oscillator model, a fourth-order contribution to the lattice energy of the face-centered cubic configuration. Further, Kestner and Sinanoğlu (1963), again using third-order perturbation theory, determined effective London dispersion forces between two molecules in a polar or nonpolar medium. In the same approximation, an important special case arises if the third body in three-atom interactions is very large (a solid), in which case one is led to an effective long-range potential between two atoms absorbed on a solid surface (Sinanoğlu and Pitzer, 1960).

All these calculations refer to the *long-range part* of a many-atom potential; consequently, they cannot be used in stability analyses. We mention that at still larger distances (those comparable to the wavelength of visible light) retardation forces take over; McLachlan (1963) has shown that the three-atom component of such forces goes over into the triple-dipole effect at much shorter distances between the atoms (still neglecting exchange).

We now turn to *short-range three-atom interactions*; these appear in a perturbation analysis if the zeroth-order wave function is taken as an *antisymmetric* product of atomic wave functions, as required by the Pauli principle, instead of a simple product. Margenau (1939) pointed out in 1939 that three-atom interactions of exchange type occur already in *first* order of perturbation theory, where they constitute a correction to the pair-repulsions between atoms at short distances. Rosen (1953) evaluated three-body exchange interactions in first order of perturbation theory between helium atoms, using a valence bond description; he considered numerically an equilateral triangle and a linear symmetric array of atoms. Shostak (1955) repeated these calculations on the basis of the molecular orbital method; he gave numerical results only for a linear symmetric array of helium atoms. The three-atom component was found to be negligible compared with pair-interactions even at the density of solid helium.

In spite of the smallness of the three-atom first-order interactions between helium atoms, it is very interesting to consider their properties in some detail. Let us denote the sum of first-order pair-interactions between a triplet of helium atoms by $E_1^{(0)}$ and the total first-order energy by E_1. Rosen finds that the relative first-order three-atom interaction, $\Delta E_1/E_1^{(0)}$, can be represented by $A \exp - a(R_{ab} + R_{ac} + R_{bc})$ both in the case of an equilateral triangle and for a linear symmetric array of helium atoms. The values of the parameters A and a are $A = -1.15$ and $a = 0.33$ for the equilateral triangle and $A = +9.8$ and $a = 0.66$ for the linear configuration (all distances expressed in atomic units). We draw the following conclusions:

(*i*) $\Delta E_1/E_1^{(0)}$ is *negative* for an equilateral triangle of atoms ($\gamma = 60°$);
(*ii*) it is *positive* for a linear symmetric array of atoms ($\gamma = 180°$);
(*iii*) its magnitude at $\gamma = 60°$ is larger than at $\gamma = 180°$ (as long as $R > 1.3$ a.u. $= 0.66$ Å).

These are *precisely the same* properties as we established earlier for the relative triple-dipole effect $\Delta E_3/E_2^{(0)}$! It appears tempting to assume that we here have to do with *general* properties of three-atom interactions, irrespective of the order of perturbation theory in which they are evaluated. Further, although no numerical results are known for $\Delta E_1/E_1^{(0)}$ for γ values other than 60° and 180°, we expect the relative first-order three-atom interactions to be more structure-sensitive than the triple-dipole effect.

Since $E_1^{(0)}$ is always positive (repulsive), Rosen's results imply that three-body first-order interactions between helium atoms *weaken* the repulsions, compared with an additive sum over pairs, for an equilateral triangle; for a linear array the three-body component *increases* the repulsion between the atoms. For the stability of solid helium (He^4) such interactions are of no consequence, since the nearest-neighbor distance is relatively very large ($R = 3.6$ Å $= 7.1$ a.u.).

If, on the other hand, we could compress solid helium so much that the nearest-neighbor distance R becomes, say, 1 Å = 1.98 a.u., then $\Delta E_1/E_1^{(0)}$ would increase to -0.18 at $\gamma = 60°$ and to $+0.066$ at $\gamma = 180°$; i.e., the three-body interactions would become very considerable in magnitude. Now, looking back at the triple-dipole effect, Eq. (20), we see that the different rare gas solids are distinguished solely by different values of the dimensionless parameter α/R^3, the cube root of which is a measure for *the dimension of the atom relative to the nearest-neighbor distance* in the lattice. *If* a similar result applies also for $\Delta E_1/E_1^{(0)}$, *then* first-order three-atom interactions may be appreciable for heavy rare gas solids.

To investigate this, we must extend the Rosen–Shostak analysis to heavy rare gas atoms. A direct extension for such many-electron systems, using explicit analytic electron wave functions, is much too complex a problem to offer a real chance of success. We will find later that a simplified analysis for heavy atoms may be carried out which incorporates the possible important parameters of the stability problem while avoiding excessive complications of many-electron systems.

So far, we have not considered three-atom *exchange* corrections to the *London–van der Waals forces* (dispersion forces). Such effects are potentially very important, since the nearest-neighbor distance in rare gas solids is practically the same as the distance of maximum attraction in the pair-potential function. At this distance, the attractive part of the pair-interaction (3) or (4) is approximately twice the repulsion for the experimental values of a or s. Consequently, three-atom interactions in *second order* of perturbation theory might have a decisive influence on the stability of rare gas crystals.

The evaluation of three-atom exchange interactions of second order is, however, very complicated indeed; not even *pair*-exchange interactions are known for any system of atoms. *First* of all, we must use an antisymmetric wave function for the triplet of atoms considered. *Second*, we cannot employ a multipole expansion for the interactions, in view of the close approach of neighbors in the crystals. *Third*, the expressions to be evaluated are of the form given by Eq. (16), which has to be summed over all excited states of the system and in which the perturbation Hamiltonian occurs twice in succession. This, in addition to the fact that we are mainly interested in *heavy* rare gas atoms which have many electrons, renders the problem of evaluating three-atom second-order exchange interactions much too formidable to handle without drastic simplifications.

McGinnies and Jansen (1956) have found that explicit calculations of the exchange effect are possible for triplets of the lightest atoms (hydrogen and helium) and for the special case where exchange must be taken into account only between *two of the three* atoms. This situation would apply to triplets of atoms of which two are nearest neighbors in the solid, whereas the third atom is considerably farther away.

Let us consider such a triplet abc of atoms, and let c be so far apart that exchange between the pair ab and the atom c can be neglected. The total second-order energy, E_2, can *in this case* be written as

$$E_2 = (E_2^{(0)})_{ab} + (E_2^{(0)})_{(ab)c}, \tag{21}$$

where the first term on the right stands for the second-order interaction for the isolated pair *ab*, whereas the second term represents the second-order interaction between the *group ab* of atoms and the atom *c*. On the other hand, the sum of pair interactions is

$$E_2^{(0)} = (E_2^{(0)})_{ab} + (E_2^{(0)})_{ac} + (E_2^{(0)})_{bc}, \tag{22}$$

so that the three-body component is given by

$$\Delta E_2 = E_2 - E_2^{(0)} = (E_2^{(0)})_{(ab)c} - (E_2^{(0)})_{ac} - (E_2^{(0)})_{bc}. \tag{23}$$

This result implies that for the evaluation of ΔE_2 we may restrict ourselves to the *partial* perturbation Hamiltonian

$$H'_{(ab)c} = H'_{ac} + H'_{bc};$$

i.e., the perturbation H'_{ab} between atoms *a* and *b* does not occur. In view of this peculiarity we may, for atom *c* sufficiently away from the group *ab*, expand $H'_{(ab)c}$ in a multipole series. If, in addition, we use the Unsöld–Vinti summation over the excited states κ of the system, then we have to evaluate only the expectation value of the square of $H'_{(ab)c}$ for a ground-state wave function which is antisymmetric with respect to exchange between *a* and *b*, multiplied by the ground-state wave function of atom *c*.

It is relatively easy for hydrogen atoms (three spins parallel) and helium atoms (using simple exponential atomic functions) to evaluate ΔE_2. Again, it is more accurate to compute the *relative* second-order effect $\Delta E_2/E_2^{(0)}$ which avoids the necessity of determining the average excitation energy E_{av} resulting from the summation over excited states.

Numerical results were computed for a rectangular configuration of atoms and for a linear symmetric array, under the condition $R_{ac} > R_{ab}$, for different values of R_{ac} and R_{ab}. It appeared that $\Delta E_2/E_2^{(0)}$ is always *negative*, except for very small values of R_{ab}, implying that the three-atom exchange interactions *decrease the attraction* compared with an additive sum over pairs for dipole-dipole interactions. In addition, the second-order contribution is at least an order of magnitude more important than the triple-dipole effect.

The results for hydrogen and helium must now be extended to the heavy rare gas atoms, but again we are faced with the difficulty of treating many-electron systems. Jansen and McGinnies (1956) have carried out such an extension on the basis of an approximation of "*effective electrons*," one such electron per atom, with spins parallel. One effective electron represents an average electron charge distribution of the atom. That such an approximation may be applied is based on explicit calculations of second-order interactions between neon atoms, using wave functions by Brown (1933) as modified by Bleick and Mayer (1934). We will postpone a detailed discussion of this approximation method until Section III, since our final calculations of crystal stability will be based on a similar method.

The analysis was limited once again to the special case of two overlapping atoms *a* and *b*, while the third atom *c* is at a larger distance from the two. In that case, as discussed above, we can employ a multipole expansion. Results were given numerically for dipole-dipole and for dipole-quadrupole interactions between the pair *ab* and the atom *c* and summed over all triplet configurations of atoms in the hcp and fcc structures with atom *b* nearest neighbor of *a* and with atom *c* within a distance $R\sqrt{15}$ from the origin, where *R* is the nearest-neighbor distance. Triplets for which *c* is a nearest neighbor of *a* or *b* were excluded; the summation region contains approximately 3600 triangles, of which 141 are different.

The results are dramatic in that they appear to depend critically upon the distance of closest approach of atom *c*. If *all* triplets are included, also those for which atom *c* is nearest neighbor to *a* or *b*, then the result is a very large repulsion. Further, the total three-body interaction changes its sign for small variations in the minimum distance between *c* and the origin at about the position of next-nearest neighbors. This reveals very clearly that stability calculations cannot be based on multipole expansions, and that exchange between all three atoms of the triplet must be taken into account at the outset, in order to arrive at reliable conclusions regarding crystal stability.

Of the many types of three-atom interactions which we have reviewed in this section, the Axilrod–Teller triple-dipole effect and the Rosen–Shostak first-order calculations for helium atoms have, in particular, provided us with important preliminary insight into the structure of the three-body potential. Of these, the first effect refers to the long-range part of the potential (no exchange), whereas the first-order interactions play the most important role for very small separations between the three atoms. We have found that these two contributions, *relative* to the sums of second- and first-order pair-interactions, surprisingly enough, have qualitatively the *same properties*.

None of the effects discussed in this section can, however, be used as a key to the explanation for the stability of rare gas crystals because of one, or more, of the following defects:

(1) They concern systems of little direct interest (helium).
(2) Overlap between charge distributions is either neglected or treated classically.
(3) A multipole expansion is employed for the interactions.

Instead, we must attempt to develop an approximation method which retains the possibly important parameters of the stability problem, which is sufficiently simple for numerical calculations, and which avoids the defects mentioned above. Such an attempt will be undertaken in Section III.

B. Many-Body Interactions between Alkali Halide Ions

In Section I,C we reviewed the stability problem of alkali halide crystals. Since the Born–Mayer theory of ionic solids has been very successful in interpreting many properties of alkali halide crystals, we felt that a solution to the stability problem must be achieved without any drastic revision of the principles underlying this model.

In order to retain the essential simplicity of the Born–Mayer theory, we must explain the dependence of the repulsive parameters B and ρ of Eq. (13) on the crystal structure, as required by the phenomenological analysis of Tosi and Fumi (1962). It was concluded that three-ion interactions are most likely to provide this explanation. We recall that the Born–Mayer energy difference between the cesium chloride and the sodium chloride configurations for heavy alkali halides is of the order of a few kilocalories per mole, i.e., of the order of 10^4 times that between the pair energies of the hcp and fcc structures for rare gas solids. Consequently, the three-ion interactions in these structures must be much larger than the triplet interactions between rare gas atoms. Further, these interactions must depend very sensitively upon crystal structure; i.e., they are of short range (exchange type). This, in turn, excludes the consideration of three-ion interactions which depend directly upon the net charge of an ion, since such effects are of long range.

The first explicit analysis of many-ion interactions in alkali halide crystals was given by Löwdin (1948, 1956). He undertook a calculation of the *total* cohesive energy for a number of alkali halides from *first* order of perturbation theory. The lattice energy was found to contain a considerable many-ion exchange component of which the first term constitutes a three-ion exchange interaction. This three-body component is very large, namely, of the order of 10 to 20 kcal/mole, with *negative* sign, for a solid consisting of small cations and large anions. The three-ion interaction is practically proportional to the Madelung constant of the crystal; its magnitude decreases rapidly if cation and anion approach equal size.

The fact that this three-ion interaction is proportional to the Madelung constant of the lattice implies that its *structure sensitivity is too low* to explain the stability of the cesium chloride configuration for the heavy cesium salts. Since the Madelung constants of the two configurations differ by only 1%, the maximum difference in three-ion interactions between these structures is only of the order of 0.1 kcal/mole. In addition, the ions of heavy cesium halides are not very different in size, so that the three-ion energy is even smaller than this number.

Löwdin applied the results of his detailed analysis also to a calculation of transition pressures for different alkali halides and to an evaluation of elastic constants for these solids. The analysis of elastic coefficients is important, since the occurrence of many-ion interactions should manifest itself in a deviation from the Cauchy relations between these coefficients. Although Löwdin did not consider second-order three-ion interactions, he believed that a combination of first- and second-order many-body potentials would probably provide an explanation for the stability of the cesium chloride configuration for the heavy cesium halides (Löwdin, 1948, p.105).

Later attempts to introduce many-ion interactions in the lattice energy of alkali halide crystals have mostly been directed towards an interpretation of Löwdin's results on a simpler basis. We mention, in particular, the so-called "effective exchange charge" or "shell" model developed by Dick and Overhauser (1958). These authors considered first-order interactions between closed shells of electrons, starting from helium atoms. The first-order energy expression for two atoms or ions with closed shells can, in part, be interpreted in terms of a positive "exchange charge distribution" which has a maximum in the region of maximum overlap between the two atoms. The exchange charge arises because of spin-correlation (Pauli principle); as a result, there is a net positive charge between the atoms and a net negative charge on the atoms. A similar model was developed independently by Hanlon and Lawson (1959). As a result of the exchange charges, a pair of overlapping atoms acquires "exchange" electric multipole moments; such moments on *different* pairs of atoms or ions interact, which gives rise to *four-atom* or *four-ion* interactions. Also, a multipole moment of one pair of ions interacts electrostatically with the net ion charge of a third ion, giving rise to *three-ion* components of the lattice energy.

To illustrate this idea, let us give a simple example (Jansen, 1962). Consider two identical atoms, each with one electron, spins parallel, at interatomic distance R. The charge distribution of this pair has cylindrical symmetry about an axis z, coinciding with the direction of R. For cylindrically symmetric distributions of charge we can define scalar multipole moments of arbitary (positive) order (Jansen, 1957, 1958). The first non-vanishing moment is in this case a quadrupole, defined by

$$Q = \langle \textstyle\sum_i e_i z_i^2 \rangle - \langle \textstyle\sum_i e_i x_i^2 \rangle, \tag{24}$$

where the x axis is perpendicular to z, but otherwise arbitrary. We denote the two atoms by a and b, the two electrons by 1 and 2. The zero-order wave function is (Slater determinant)

$$\Psi_0 = \{\varphi_a(1)\varphi_b(2) - \varphi_a(2)\varphi_b(1)\}/[2(1-\Delta^2)^{1/2}], \tag{25}$$

where Δ is the overlap integral, $\Delta = \int \varphi_a\varphi_b\, d\tau$. (Spin functions may be omitted, since the perturbation Hamiltonian is spin-independent.) The direct terms of Ψ_0^2 give, of course, no contribution to Q or to any other permanent multipole moment, so that we are left with the exchange term $-2\varphi_a(1)\varphi_b(1)\varphi_a(2)\varphi_b(2)$.

We denote the z coordinates of the position vectors of electrons 1 and 2, with respect to the center of the line connecting the two nuclei, by z_1 and z_2, respectively. Then we obtain

$$\langle \textstyle\sum_i e_i z_i^2 \rangle \equiv [4eI_3\Delta - 4e(R/2)^2\Delta^2]/2(1-\Delta^2), \tag{26}$$

where

$$I_3 = \int z^2 \varphi_a\varphi_b\, d\tau$$

and

$$\langle \textstyle\sum_i e_i x_i^2 \rangle = 4eI_2\Delta/2(1-\Delta^2)$$

with

$$I_2 = \int x^2 \varphi_a\varphi_b\, d\tau. \tag{27}$$

The expression for the scalar quadrupole moment of the pair ab of atoms then becomes, with Eqs. (24), (26), and (27),

$$\begin{aligned} Q &= \langle \textstyle\sum_i e_i z_i^2 \rangle - \langle \textstyle\sum_i e_i x_i^2 \rangle \\ &= [4e\Delta(I_3 - I_2) - 4e\Delta^2(R/2)^2]/2(1-\Delta^2). \end{aligned} \tag{28}$$

It is easiest to calculate I_2 and I_3 for a Gaussian distribution of charge of the electrons 1 and 2 on atoms a and b. In that case $I_2 = I_3$, so that

$$Q = -2e[\Delta^2/(1-\Delta^2)](R/2)^2. \tag{29}$$

This quadrupole moment is the same as that caused by an effective negative point charge (*exchange charge*)

$$-(\delta e) = -e[\Delta^2/(1-\Delta^2)]$$

at nucleus a, the same charge at nucleus b, and a positive charge $2(\delta e)$ at the center between the two nuclei.

In addition to a quadrupole moment, exchange multipoles of higher (even) order occur. If we now consider first-order interactions between two pairs of atoms at *large distances* (the overlap between the two pairs must be negligible), then the interaction contains contributions due to the exchange multipole moments of the two pairs. In crystals of high symmetry the contribution of such interactions *to the lattice energy* may be neglected altogether.

Dick and Overhauser (1958) showed that for alkali halide crystals this model leads to a generalization of the Clausius-Mossotti, Lorenz-Lorentz, and Szigeti equations of dielectric theory. Dick (1963) used the same method later for a calculation of elastic constants for alkali halide crystals, but the limits of validity of the exchange charge model render such an application very doubtful. An extension was suggested by Cochran (1959) and by Woods *et al.* (1960) in a study of lattice vibrations of germanium and of alkali halides. For the stability problem of alkali halide crystals this type of consideration is clearly not applicable.

Finally, we mention that Colwell (1960) has attempted a detailed analysis for the specific case of cesium chloride (together with some estimates for potassium chloride) in terms of the various possible contributions to the crystal field, starting from explicit expressions for the electron wave functions. Such a *global* method is unsuited to discern explicitly many-ion contributions to the crystal energy, and in view of the many approximations involved it is impossible to estimate its reliability.

To our knowledge, no other attempts have been made in the literature to analyze possible many-ion contributions to the lattice energy of alkali halides. In principle, the triple-dipole effect and the Rosen–Shostak first-order three-body interactions should occur also with alkali halides, although the triple-dipole interactions are clearly negligible, on a relative scale, in this case. Further, although van der Waals forces are not explicitly represented in the simple Born–Mayer potential, three-ion exchange components of these forces must explicitly be considered for stability.

So far, then, we have not found a promising starting point for the analysis of three-ion interactions in alkali halide crystals and their effect on the stability of these solids, except in the analogy with rare gas crystals. We will, therefore, in the next section combine the two stability problems and analyze them *on the same* basis.

III. Stability of Rare Gas and Alkali Halide Crystals in Terms of Three-Body Interactions

A. General. The Effective-Electron Approximation

The review of the stability problems for rare gas and alkali halide crystals which we have given in the previous sections leads to the following conclusions.

(*i*) The stability of the fcc configuration for rare gas crystals, as well as that of the CsCl configuration for the heavier cesium chlorides, is essentially due to the effect of three-atom and three-ion interactions in these solids.

(*ii*) The three-body interactions must be considerable, i.e., of low orders of perturbation theory, and they must be very sensitive with respect to crystal structure; i.e., they are of short range (exchange type).

(*iii*) Since the properties of three-body interactions in the two cases are qualitatively the same, we assume that they have the same physical origin. This similarity is further based on the fact that the ions of alkali halides are isoelectronic with rare gas atoms and that both consist of closed electron shells. Consequently, their interactions must be of the same form if we subtract purely electrostatic contributions, due to the net ion charges, to the lattice energy of alkali halides and disregard polarization effects in view of the high symmetry of the unstrained crystals.

It is clear that explicit calculations with detailed wave functions for each rare gas solid and each alkali halide crystal individually would be much too complex to offer any chance of meaningful results. In addition, it is not necessary to go to such details. To see this, we note that stability of one, or the other, of the crystal structures under consideration is an experimental phenomenon *common to a number of representatives of both the rare gas atoms and the alkali halides.* Consequently, stability cannot depend sensitively upon the detailed analytic form of the wave functions, but it must be determined by more general parameters characterizing the charge distributions of the atoms or ions.

In order to find these parameters, we carried through explicit calculations of second-order exchange contributions to the energy of interaction between neon atoms (Jansen and McGinnies, 1956), using Bleick–Mayer wave functions (Bleick and Mayer, 1934). On the basis of the Unsöld–Vinti summation over excited states of the system, we have to evaluate the expectation value of the square of the perturbation Hamiltonian for the ground-state wave function. Such an expression can formally be written as a series of terms of ascending order of electron exchange between two neon atoms. The first term gives the direct contribution to the second-order interactions (no exchange); in the second term the wave functions to the left and to the right of H'^2 in the expectation value differ only by exchange of *one* pair of electrons between the two atoms (single exchange); in the third term these components of the wave function differ in the coordinates of two pairs of electrons (double exchange); and so forth.

We have compared single- and double-exchange effects for neon atoms and found that double exchange is of an order of magnitude less important than single exchange, up to densities of solid neon (nearest-neighbor distance $R = 3.2$ Å). In addition, each order of *inter*atomic exchange is associated with *intra*-atomic exchange of electrons to various orders (coupled exchange). We found that coupled exchange effects need not be considered in solid neon. Since neon crystallizes already in the cubic configuration, *it follows that also for the other rare gas crystals we may neglect effects due to multiple and coupled exchange.*

We are then left with an average of *single-exchange* effects, each term involving one electron of each atom (ion) of a pair. This average may be replaced by an exchange between "effective" electrons, one such electron per atom (ion), with spins parallel, each electron representing an average of the charge distribution of that atom or ion. Further, we recall that the interactions, with which we are here concerned, are *very weak* relative to the magnitude of the ionization potentials of the atoms or ions. (As an example, we noted that the interaction between two nearest neighbors in solid xenon is approximately seventy times weaker than the energy of the xenon-fluorine bond in XeF_2.) Now, in the case of such extremely weak interactions an atom in the field of force of a second atom experiences only the effect of an average charge distribution of the first atom; i.e., the detailed electronic structure of the atom plays no role. Therefore, in the charge distribution of an effective electron only the "size" of the atom can enter as a parameter.

The problem of three interacting atoms or ions becomes then formally the same as that for three hydrogen atoms with parallel spins of the electrons. In principle, to represent also ions with charges plus one and minus one, we should use *two* effective electrons per ion, counterbalanced by nuclear charges of plus three and plus one, respectively. The exchange effect then becomes formally the same as that between ions of lithium hydride. However, we subtract electrostatic interactions between the net charges from the outset, so that ions and atoms can be treated on the same basis.

The charge distribution of an effective electron is as yet undetermined. Since, however, stability depends critically upon the interactions between atoms (or ions), we choose this distribution in such a way that it represents as accurately as possible the pair-potential function between neon atoms. This appears possible if we select a Gaussian charge distribution

$$\rho(r) = (\beta/\pi^{1/2})^3 \exp(-\beta^2 r^2), \tag{30}$$

where r is the distance from the effective electron to its nucleus and where β measures the width of the distribution. Let us first determine β in such a way that it fits the R^{-6} part of the pair-potential between rare gas atoms. We consider two such Gaussian atoms at *large* distance R. The London dipole-dipole interactions are given by

$$E_2(\text{dip-dip}) = -e^4/4\beta^4R^6E_{\text{av}}, \tag{31}$$

where E_{av} is an average excitation energy for the specific atom considered, which we will assume to be approximately the same as the first ionization potential. This expression is now put equal to the inverse-six term of the Lennard-Jones (12, 6) potential, i.e.,

$$3e^4/4\beta^4R^6E_{\text{av}} = 4\varepsilon\sigma^6/R^6,$$

from which it follows that

$$\beta^4 = 3e^4/16\varepsilon\sigma^6E_{\text{av}}. \tag{32}$$

The values of β can, therefore, be determined by using empirical values for ε, σ, and E_{av}. We note, however, that the precise values of β do not matter, since we are mainly interested in a *range* of such values for heavy rare gas atoms. For neon, argon, krypton, and xenon Eq. (32) yields values for β equal to 1.07, 0.623, 0.532, and 0.454, in units 10^8 cm^{-1}.

Next, we place the same Gaussian atoms at *short* distances from each other and evaluate the repulsion in first order of perturbation theory. The zeroth-order wave function is an antisymmetric product of wave functions for the two Gaussian atoms, taken as the positive square root of Eq. (30). This problem can easily be solved with a charge distribution (30), owing to the computational simplicity of integrals containing Gaussian functions (for details, we refer to Section III,B). The result, as a function of β and of the distance R, is now put equal to the repulsive part of a Lennard-Jones (12, 6) potential or, more accurately, to the repulsive part of a modified (exp-six) Buckingham potential as given by Eq. (4). The results show (Jansen, 1962) *excellent* agreement for neon atoms; i.e., the *same* value of β fits the long-range and the short-range parts of the pair-potential accurately. For the heavier rare gas atoms β appears to *increase* somewhat with decreasing interatomic distances. Since only approximate values matter, we will retain the β values determined from the long-range part of the potential.

For alkali halides such a procedure cannot be followed, since pair-potentials between the ions are not known with comparable accuracy. Since the alkali halide ions are isoelectronic with rare gas atoms, we expect that the β value for each cation will be somewhat larger than that for the corresponding isoelectronic rare gas atom, and that for each anion it is somewhat smaller than this value.

To estimate such *differences*, experimental values for the molar diamagnetic susceptibilities of alkali halide ions and rare gas atoms (Kittel, 1957, Chapter 8) were used to evaluate β parameters of the ions (Lombardi and Jansen, 1964a). Since the susceptibility, χ, of an atom or ion with one effective electron is proportional to $\langle r^2 \rangle$, i.e., proportional to β^{-2} for a Gaussian distribution, we have the following relation:

$$(\beta_{\text{ion}}/\beta_{\text{atom}})^2 = \chi_{\text{atom}}/\chi_{\text{ion}}. \tag{33}$$

The β values for ions were calculated from Eq. (33) by comparing each ion with its corresponding isoelectronic rare gas atom, using for the atoms the β parameters obtained from pair-potentials. It appears that for rare gas solids the three-atom interactions are uniquely determined by the dimensionless quantity βR(R is the nearest-neighbor distance in the lattice) and by the type of structure (hcp or fcc). For alkali halides the relevant parameters are βR, where β denotes either cation or anion, the crystal structure (NaCl or CsCl), and the ratio of β values for cation and anion. We adopt the convention that β will denote the *larger* ion and β' the *smaller* ion and we introduce the parameter $\gamma = (\beta'/\beta)^2 > 1$ as a measure for their relative size. In most cases, then, β represents the anion and β' the cation, except if the cation is very large (CsF, for example).

In Table I we give the resulting β and β' values for alkali halides together with those of rare gas atoms. In Table II, the values of βR and γ are listed for the alkali halide crystals; β represents the larger ion. For rare gas crystals the values of the dimensionless parameter βR are 3.4 (solid neon), 2.4 (solid argon), 2.1 (solid krypton), and 2.0 (solid xenon).

TABLE I

VALUES OBTAINED WITH EQ. (33) FOR GAUSSIAN PARAMETER β, IN UNITS 10^8 CM^{-1}, FOR ALKALI HALIDE IONS, AND COMPARED WITH THOSE OF RARE GAS ATOMS

He	1.59	Ne	1.07	A	0.623	Kr	0.532	Xe	0.454
Li^+	2.60	Na^+	1.162	K^+	0.718	Rb^+	0.600	Cs^+	0.503
		F^-	0.936	Cl^-	0.558	Br^-	0.479	I^-	0.419

The values in the two tables should be viewed primarily in relation to each other and to the rare gas crystals. We have included the calculated parameters for lithium halides, although it is only relevant to mention that their γ values are high.

TABLE II

VALUES OBTAINED WITH EQ. (33) FOR THE DIMENSIONLESS PARAMETERS βR AND $\gamma = (\beta'/\beta)^2$ FOR ALKALI HALIDE CRYSTALS[a]

Alkali halide	βR	γ	Alkali halide	βR	γ
LiF	1.88	7.72	LiBr	1.325	29
NaF	2.16	1.54	NaBr	1.43	5.88
KF	1.91	1.70	KBr	1.58	2.25
RbF	1.69	2.43	RbBr	1.64	1.57
CsF	1.51	3.46	CsBr	1.78	1.10
LiCl	1.44	22	LiI	1.26	38
NaCl	1.57	4.34	NaI	1.35	7.69
KCl	1.75	1.66	KI	1.48	2.94
RbCl	1.83	1.16	RbI	1.54	2.05
CsCl	1.79	1.23	CsI	1.66	1.44

[a] R denotes the nearest-neighbor distance and β represents the larger ion.

After this simplification of the stability problems, first- and second-order perturbation calculations are carried out for triplets of atoms (ions), and the results are summed over the respective lattices. Since the three-atom and three-ion interactions are of exchange type and, therefore, of very short range, we will limit ourselves to triplets formed by a central atom and any two of its *nearest* neighbors. In the case of alkali halides it generally proves necessary to include also triplets formed by a central ion and any two of its *next-nearest* neighbors.

B. Rare Gas Atoms. First- and Second-Order Three-Atom Exchange Interactions

We will now evaluate first- and second-order exchange interactions for a triplet (abc) of rare gas atoms; the electron charge distribution of each atom is replaced by that of an effective electron with characteristic parameter β. By "exchange interactions" we mean that exchange of electrons between the different atoms is taken into account for their evaluation, *not* that these interactions are solely due to exchange effects. The methods and results reported in this section correspond to those published recently (Jansen, 1962, 1963a,b, 1964).

For the atomic wave functions we take

$$\varphi(r) = \rho^{1/2}(r) = (\beta/\pi^{1/2})^{3/2} \exp(-\beta^2 r^2/2); \tag{30a}$$

the zero-order wave function for the triplet, with three electron spins parallel, is then (Slater determinant):

$$\Psi_0 = [3!(1 - \Delta^2_{abc})]^{-1/2} \det\{\varphi_a(1)\varphi_b(2)\varphi_c(3)\}, \tag{34}$$

where a, b, and c denote the atoms, 1, 2, and 3 the effective electrons, and where

$$\Delta^2_{abc} = \Delta^2_{ab} + \Delta^2_{ac} + \Delta^2_{bc} - 2\Delta_{ab}\Delta_{ac}\Delta_{bc} \tag{35}$$

in terms of the overlap integrals Δ_{ab}, etc., between the different pairs of atoms. The perturbation Hamiltonian, H'_{abc}, can be written as

$$H'_{abc} = H'_{ab} + H'_{ac} + H'_{bc}$$

in terms of the perturbations between the pairs. An important advantage of the Gaussian effective-electron approximation is that it enables us to evaluate all three-center integrals, both in first and in second orders of perturbation theory, *without* making use of multipole expansions. In the previous sections we have seen that this is an essential requirement for the results to be meaningful.

1. *First-Order Calculations*

Let $E_1 = \langle H'_{abc} \rangle$ denote the *total* first-order energy for the (arbitrary) triplet *abc*, where the expectation value is determined with the zero-order wave function (34), and let $E_1^{(0)} = \langle H'_{abc} \rangle^{(0)}$ denote the sum of first-order interactions between the three *isolated* pairs of atoms which form the triangle with the corresponding zero-order pair wave functions. We will determine the relative three-body component of the first-order energy, i.e.,

$$\Delta E_1/E_1^{(0)} = (E_1 - E_1^{(0)})/E_1^{(0)}, \tag{36}$$

as a function of the Gaussian parameter β and of the triangular dimensions. Now $\langle H'_{abc} \rangle = \langle H'_{ab} \rangle + \langle H'_{ac} \rangle + \langle H'_{bc} \rangle$ and $\langle H'_{abc} \rangle^{(0)} = \langle H'_{ab} \rangle^{(0)} + \langle H'_{ac} \rangle^{(0)} + \langle H'_{bc} \rangle^{(0)}$, so that we can write

$$E_1 - E_1^{(0)} = \{\langle H'_{ab} \rangle - \langle H'_{ab} \rangle^{(0)}\} + \{\langle H'_{ac} \rangle - \langle H'_{ac} \rangle^{(0)}\} + \{\langle H'_{bc} \rangle - \langle H'_{bc} \rangle^{(0)}\}.$$

In other words, to determine $\Delta E_1/E_1^{(0)}$ it is sufficient to calculate $\{\langle H'_{ab} \rangle - \langle H'_{ab} \rangle^{(0)}\}$ and $\langle H'_{ab} \rangle^{(0)}$; the rest follows by permuting indices.

After substitution of H'_{ab} and of the expression (34) for Ψ_0 into $\langle H'_{abc}\rangle$, we obtain the following result:

$$\begin{aligned}\langle H'_{ab}\rangle/e^2 = {} & \frac{1}{R_{ab}} - 2\,\frac{1-\frac{1}{2}(\Delta_{ac}^2+\Delta_{bc}^2)}{1-\Delta_{abc}^2}\,G_{aa(b)} \\ & + 2\,\frac{\Delta_{ab}-\Delta_{ac}\Delta_{bc}}{1-\Delta_{abc}^2}\,G_{ab(a)} + \frac{\Delta_{ac}-\Delta_{ab}\Delta_{bc}}{1-\Delta_{abc}^2}\,G_{ac(b)} \\ & + \frac{\Delta_{bc}-\Delta_{ab}\Delta_{ac}}{1-\Delta_{abc}^2}\,G_{bc(a)} + \frac{1}{1-\Delta_{abc}^2}\,(A_{abab}-A_{aabb}) \\ & + \left(\frac{\Delta_{ac}}{1-\Delta_{abc}^2}\right)(A_{abbc}-A_{abcb}) + \left(\frac{\Delta_{bc}}{1-\Delta_{abc}^2}\right)(A_{aabc}-A_{abac}), \qquad (37)\end{aligned}$$

where the symbols G and A are abbreviations for the following integrals:

$$G_{ab(c)} = \int \frac{\varphi_a\varphi_b}{r_c}\,d\tau,$$

$$G_{aa(b)} = \int \frac{\varphi_a\varphi_a}{r_b}\,d\tau,$$

etc., with r_c = distance between an electron and nucleus c and r_b = distance between an electron and nucleus b, and

$$A_{abac} = \iint \frac{\varphi_a(1)\varphi_b(2)\varphi_a(1)\varphi_c(2)}{r_{12}}\,d\tau_1\,d\tau_2,$$

$$A_{aabc} = \iint \frac{\varphi_a(1)\varphi_a(2)\varphi_b(1)\varphi_c(2)}{r_{12}}\,d\tau_1\,d\tau_2,$$

etc., with r_{12} = distance between electrons 1 and 2.

On the other hand, the first-order energy between a and b, *without* c present, is obtained from Eq. (37) by removing atom c to infinity. The result is:

$$\langle H'_{ab}\rangle^{(0)}/e^2 = \frac{1}{R_{ab}} - \frac{1}{1-\Delta_{ab}^2}\,[2G_{aa(b)} - 2\Delta_{ab}G_{ab(a)} - (A_{abab}-A_{aabb})]. \qquad (38)$$

The three-body component of $\langle H'_{ab}\rangle/e^2$, namely

$$\{\langle H'_{ab}\rangle - \langle H'_{ab}\rangle^{(0)}\}/e^2,$$

is obtained by subtracting Eq. (38) from Eq. (37). We do not need any further equations, since $\langle H'_{ac}\rangle$, for example, is obtained from Eq. (37) by replacing b by c.

The integrals G and A, for Gaussian distributions of charge, have been evaluated by Boys (1950). It is convenient to use the following abbreviation:

$$F(x^2) \equiv \frac{1}{x}\int_0^x \exp(-y^2)\,dy = (\pi^{1/2}/2)\,\frac{\operatorname{erf} x}{x};$$

the integrals G and A can then be expressed in terms of the F function and of overlap integrals. The latter can also be evaluated easily, since, e.g.,

$$\Delta_{ab}^2 = \exp(-\beta^2 R_{ab}^2/2)$$

for Gaussian functions; R_{ab} is the internuclear distance between atoms a and b. It is most convenient to determine F and Δ by interpolation, using *Tables of the Error Function and Its Derivative* (National Bureau of Standards, Washington, D.C., 1935). For solid neon, argon, krypton, and xenon the (long-range) values of βR are 3.4, 2.4, 2.1, and 2.0, respectively. We give in Table III a list of integrals and their values for Gaussian distributions, occurring in the expression for E_1.

TABLE III

LIST OF INTEGRALS FOR E_1

Integral	Value for Gaussian distribution
Δ_{ab}^2	$\exp(-\beta^2 R_{ab}^2/2)$
Δ_{ac}^2	$\exp(-\beta^2 R_{ac}^2/2)$
Δ_{bc}^2	$\exp(-\beta^2 R_{bc}^2/2)$
$\Delta_{abc}^2 = \Delta_{ab}^2 + \Delta_{ac}^2 + \Delta_{bc}^2 - 2\Delta_{ab}\Delta_{ac}\Delta_{bc}$	
$G_{aa(b)} = G_{bb(a)}$	$(2\beta/\sqrt{\pi})F(\beta^2 R_{ab}^2)$
$G_{ab(a)} = G_{ab(b)}$	$(2\beta/\sqrt{\pi})\Delta_{ab}F(\beta^2 R_{ab}^2/4)$
$G_{bc(a)}$	$(2\beta/\sqrt{\pi})\Delta_{bc}F(\beta^2 R_{a(bc)}^2)$
$G_{ac(b)}$	$(2\beta/\sqrt{\pi})\Delta_{ac}F(\beta^2 R_{b(ac)}^2)$
A_{abab}	$\beta(2/\pi)^{1/2}F(\beta^2 R_{ab}^2/2)$
A_{aabb}	$\beta(2/\pi)^{1/2}\Delta_{ab}^2$
A_{abac}	$\beta(2/\pi)^{1/2}\Delta_{bc}F(\beta^2 R_{a(bc)}^2/2)$
A_{abcb}	$\beta(2/\pi)^{1/2}\Delta_{ac}F(\beta^2 R_{b(ac)}^2/2)$
A_{aabc}	$\beta(2/\pi)^{1/2}\Delta_{ab}\Delta_{ac}F(\beta^2 R_{bc}^2/8)$
A_{abbc}	$\beta(2/\pi)^{1/2}\Delta_{ab}\Delta_{bc}F(\beta^2 R_{ac}^2/8)$

For the application to the stability problem of rare gas crystals we limit ourselves, as stated before, to triangles formed by a central atom and any two of its nearest neighbors in the crystal. There are twelve nearest neighbors in both the fcc and hcp structures, so that we have $(12 \times 11)/2 = 66$ such triangles in either lattice. Any one of these isosceles triangles is specified by the nearest-neighbor distance, R, and by the opening θ, of the triangle at the central atom. It appears that $\Delta E_1/E_1^{(0)}$ is a function only of the dimensionless parameter βR and of θ.

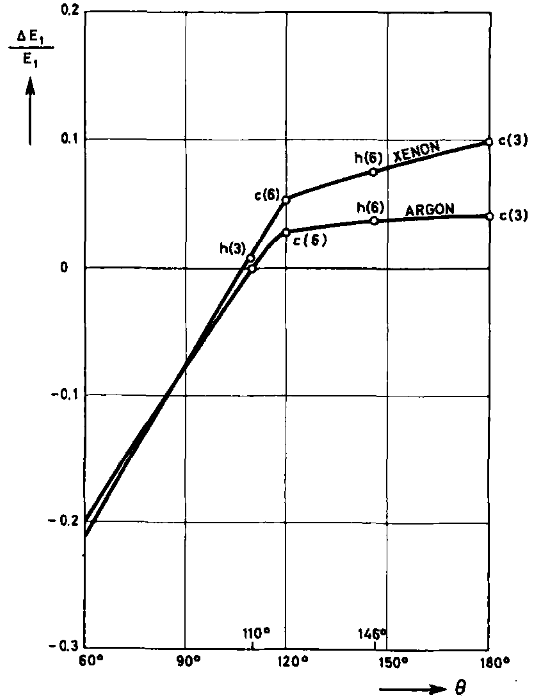


Fig. 2. Relative first-order three-body interaction $\Delta E_1/E_1^{(0)}$ for triangles $a = b = 1$, c (units of nearest-neighbor distance) of argon and xenon atoms, as a function of the angle θ between the sides a and b. The symbols h and c denote the hcp and fcc lattice, respectively; the number of triangles at a specific angle θ is given in parentheses.

The numerical results for $\Delta E_1/E_1^{(0)}$ are given in Fig. 2 for solid argon and solid xenon, as a function of the opening θ of the isosceles triangle at the central atom. The values for solid krypton lie in between those for argon and xenon; for solid neon the results are not sufficiently accurate because all three-body interactions become very small; these results indicate that the negative part of $\Delta E_1/E_1^{(0)}$ is close to that of argon or xenon, but that the positive part ($\theta > 110°$) lies close to the horizontal axis.

We note from Fig. 2 that, for values of θ between 60° and approximately 110°, $\Delta E_1/E_1^{(0)}$ is negative. Since $E_1^{(0)}$ is always positive (repulsion), this implies that three-body interactions for such triangles *weaken the interatomic repulsion* as compared with an additive sum over pairs. For θ larger than 110° the first-order interactions are *more repulsive* than the sum of interactions for the three isolated pairs. *This change of sign agrees with the results found by Rosen* (1953) [Shostak (1955) analyzed only the case $\theta = 180°$] for three helium atoms.

It is easy to show that a curve of the type as given in Fig. 1 *stabilizes the hexagonal close-packed configuration.* To see this we determine first the differences in triangular arrangements between the fcc and hcp structures; i.e., we compare the sixty-six triangles formed by a central atom and any two of its twelve nearest neighbors in the two lattices. It appears that of these sixty-six triangles fifty-seven are the same between the two structures, but nine are different. The dimensions of these nine triangles and the corresponding values of θ are given in Table IV (a, b, and c stand here not for the atoms but for the sides of the triangle in units of nearest-neighbor distance).

TABLE IV

THE NINE DIFFERENT TRIANGLES a, b, c WITH $a = b = 1$ BETWEEN THE HCP AND FCC STRUCTURES

	a^2, b^2, c^2	θ	No.	a^2, b^2, c^2	θ	No.
hcp	1, 1, $\frac{8}{3}$	109° 28′	3	1, 1, $\frac{11}{3}$	144° 44′	6
fcc	1, 1, 3	120°	6	1, 1, 4	180°	3

For a comparison with $\Delta E_1/E_1^{(0)}$ we have, in Fig. 1, also indicated the important values $\theta = 109°$ and 145° (hcp) and $\theta = 120°$ and 180° (fcc), supplied the corresponding points of the curves with h (hexagonal) and c (cubic), and written in parentheses the number of triangles for that value of θ in the corresponding structure according to Table IV.

Since the coordination number of the two structures is the same, we determine the first-order energies for the same values of βR. To compare the two lattices, the values of ΔE_1 have to be calculated for a *fixed* value of $E_1^{(0)}$ itself, for example, at $\theta = 120°$. This transformation of the total pair-interactions may be carried out according to a $1/R^{12}$ dependence, or an exponential decrease of the pair-repulsion, which makes very little difference. It flattens the curve for $\Delta E_1/E_1^{(0)}$ slightly between $\theta = 120°$ and 180°, leaves the zero value at $\theta \approx 110°$ unchanged, and lowers the value for an equilateral triangle somewhat. This, however, does not affect the following argument.

In good approximation $\Delta E_1/E_1^{(0)}$ increases (slowly and) linearly with θ between 120° and 180°. Let us indicate its value at 145° by X, at 180° by $X + \alpha$. Then at 120° the value is very nearly $X - \alpha$, whereas at $\theta = 109°$ the contribution is $Y \approx 0$. The comparison yields ($E_1^{(0)}$ evaluated at fixed θ):

$$\begin{aligned}[E_1(\text{fcc}) - E_1(\text{hcp})]/E_1^{(0)} &= 6(X - \alpha) + 3(X + \alpha) - (6X + 3Y) \\ &= 3[(X - \alpha) - Y] > 0. \qquad (39)\end{aligned}$$

Since $E_1^{(0)}$ is always positive, this means that the fcc lattice has a higher (positive) first-order energy than the hcp configuration, so that *the hexagonal close-packed lattice is stabilized by first-order three-body interactions.*

Numerically, the difference (39) is found to be approximately 10% for solid argon and about 15% for solid xenon. Relative to the *total* first-order energy of the lattice, the difference amounts to a few per cent for all heavy rare gas solids; it is, therefore, of the correct order of magnitude, but has the *wrong sign*. We see here, however, that three-atom interactions can give rise to energy differences between the hcp and fcc structures which are much larger than the 0.01% potential barrier in favor of the hcp structure because of pair-interactions.

The type of curve of Fig. 2 is of particular importance also for *second-order* three-body interactions, which we will now analyze.

2. *Second-Order Calculations*

The result of the first-order three-atom interactions as applied to rare gas crystal stability was found to be *negative* in the sense that these first-order interactions render the hexagonal close-packed structure even more stable, in contradiction with experiment. It should be noted, however, that rare gas crystals are held together by *second-order* (van der Waals) interactions. Therefore, three-body components of van der Waals forces may be of essential importance for crystal stability. This expectation is, in fact, confirmed by the following analysis (Jansen, 1963b, 1964; Jansen and Zimering, 1963).

We consider again a triangle (abc) of atoms and three effective electrons 1, 2, 3. For the second-order energy one has to evaluate

$$E_2 = \sum_{\kappa \neq 0} \frac{(H'_{abc})_{0\kappa}(H'_{abc})_{\kappa 0}}{E_0 - E_\kappa} \equiv -\frac{1}{E_{\text{av}}} \langle [H'_{abc} - \langle H'_{abc} \rangle]^2 \rangle, \qquad (40)$$

where E_{av} is an average excitation energy defined by the averaging procedure. The index κ numbers the excited states of the system (energy E_κ)

and E_0 is the unperturbed ground-state energy. The brackets denote again an expectation value for the ground-state wave function (34).

Since $\langle H'_{abc}\rangle$ can be taken directly from the first-order calculations, the only unknown quantity is $\langle H'^2_{abc}\rangle$. Let $E_2^{(0)}$ denote the sum of second-order energies between the three *isolated pairs* of atoms which form the triangle. Then the quantity

$$\Delta E_2/E_2^{(0)} = (E_2 - E_2^{(0)})/E_2^{(0)} \tag{41}$$

measures the relative second-order three-body interactions for the given triplet of atoms. It should be noted that the average excitation energies defined by the expressions for E_2 and $E_2^{(0)}$ are not necessarily the same quantities. It can be shown (Jansen and McGinnies, 1956), however, that their difference may be ignored for the present purposes. In forming, then, the ratio $\Delta E_2/E_2^{(0)}$, the quantities E_{av} *cancel* to a sufficient degree of accuracy.

The next step consists in substituting the explicit expression for H'_{abc} in Eq. (40), together with the determinantal wave function (34), and in evaluating the various expectation values which occur. We retain again H'_{abc} in its exact form; i.e., no multipole expansions are invoked for the interactions.

Whereas the resulting equation for $E_2^{(0)}$ is relatively simple, the formal expression for $\langle H'^2_{abc}\rangle$ is already of considerable complexity. The square of Ψ contains twenty-one different terms and, if we decompose H'^2_{abc} into squares and double products:

$$H'^2_{abc} = H'^2_{ab} + H'^2_{ac} + H'^2_{bc} + 2H'_{ab}H'_{ac} + 2H'_{ab}H'_{bc} + 2H'_{ac}H'_{bc},$$

then there are in total 126 integrals to be evaluated, many of three-center type.

To arrange these terms and simplify the notation, we follow a two-step procedure. First, it appears possible to express $E_2^{(0)}$ and $\langle H'^2_{abc}\rangle$ as a linear combination of eight *auxiliary functions*, $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$, $\mathscr{D}$, $\mathscr{E}$, $\mathscr{F}$, $\mathscr{G}$, and $\mathscr{H}$. It is then found that these auxiliary functions, in turn, are all linear combinations of nine *basic integrals*, which in part can be evaluated by series expansions, whereas the remaining integrals must be evaluated by electronic computation.

The auxiliary functions are defined as follows:

$$\beta^2\mathscr{A}(\beta R_{ab}) = \iint \varphi_a^2(1)\varphi_b^2(2)H'^2_{ab}\,d\tau_1\,d\tau_2; \tag{42}$$

$$\beta^2\Delta_{ab}^2\mathscr{B}(\beta R_{ab}) = \iint \varphi_a(1)\varphi_a(2)\varphi_b(1)\varphi_b(2)H'^2_{ab}\,d\tau_1\,d\tau_2; \tag{43}$$

$$\beta^2\mathscr{C}(\beta R_{ab}, \beta R_{ac}, \sphericalangle\, bac) = \iiint \varphi_a^2(1)\varphi_b^2(2)\varphi_c^2(3)H'_{ab}H'_{ac}\, d\tau_1\, d\tau_2\, d\tau_3; \quad (44)$$

$$\beta^2\Delta_{ab}^2\mathscr{D}(\beta R_{ac}, \beta R_{c(ab)}, \sphericalangle\, ac(ab))$$
$$= \iiint \varphi_a(1)\varphi_a(2)\varphi_b(1)\varphi_b(2)\varphi_c^2(3)H'^2_{ac}\, d\tau_1\, d\tau_2\, d\tau_3; \quad (45)$$

$$\beta^2\Delta_{ab}^2\mathscr{E}(\beta R_{ac}, \beta R_{bc}, \beta R_{c(ab)})$$
$$= \iiint \varphi_a(1)\varphi_a(2)\varphi_b(1)\varphi_b(2)\varphi_c^2(3)H'_{ac}H'_{bc}\, d\tau_1\, d\tau_2\, d\tau_3; \quad (46)$$

$$\beta^2\Delta_{ab}^2\mathscr{F}(\beta R_{ab}, \beta R_{ac}, \beta R_{c(ab)})$$
$$= \iiint \varphi_a(1)\varphi_a(2)\varphi_b\ 1)\varphi_b(2)\varphi_c^2(3)H'_{ab}H'_{ac}\, d\tau_1\, d\tau_2\, d\tau_3; \quad (47)$$

$$\beta^2\Delta_{ab}\Delta_{ac}\Delta_{bc}\mathscr{G}(\beta R_{ab}, \beta R_{a(bc)}, \beta R_{(ab)(bc)})$$
$$= \iiint \varphi_a(1)\varphi_a(2)\varphi_b(2)\varphi_b(3)\varphi_c(1)\varphi_c(3)H'^2_{ab}\, d\tau_1\, d\tau_2\, d\tau_3; \quad (48)$$

$$\beta^2\Delta_{ab}\Delta_{ac}\Delta_{bc}\mathscr{H}(\beta R_{ac}, \beta R_{bc}, \beta R_{b(ac)})$$
$$= \iiint \varphi_a(1)\varphi_a(2)\varphi_b(2)\varphi_b(3)\varphi_c(1)\varphi_c(3)H'_{ac}H'_{bc}\, d\tau_1\, d\tau_2\, d\tau_3. \quad (49)$$

Similar expressions are obtained by permuting a, b, and c. In the above notation $R_{c(ab)}$ denotes the distance between atom c and the middle of R_{ab}, whereas $R_{(ab)(bc)}$ stands for the distance between the middles of R_{ab} and R_{bc}. Further, $\sphericalangle bac$ denotes the angle between R_{ab} and R_{ac}, $\sphericalangle ac(ab)$ the angle between R_{ac} and $R_{c(ab)}$, etc.

We can now write the second-order pair energy, $E_2^{(0)}$, for the triplet (abc) and the expectation value $\langle H'^2_{abc}\rangle$ occurring in the total second-order energy E_2 in terms of the auxiliary functions as follows:

$$-E_{\text{av}}E_2^{(0)}/\beta^2e^4 = \frac{\mathscr{A}(\beta R_{ab}) - \Delta_{ab}^2\mathscr{B}(\beta R_{ab})}{1 - \Delta_{ab}^2} - \langle H'_{ab}\rangle^2/\beta^2e^4 + [(ac), (bc)], \quad (50)$$

where e is the electronic charge and where $[(ac), (bc)]$ signifies that the corresponding expressions for the pairs ac and bc must be added. The expression for $\langle H'_{abc}\rangle^2$ can be taken directly from the first-order calculations.

The equation for $\langle H'^2_{abc} \rangle$ reads:

$$\begin{aligned}(1-\Delta^2_{abc})\langle H'^2_{abc}\rangle/\beta^2 e^4 = {} & \{\mathscr{A}(\beta R_{ab}) - \Delta^2_{ab}\mathscr{B}(\beta R_{ab}) + 2\mathscr{C}(\beta R_{ab}, \beta R_{ac}, \sphericalangle bac) \\ & - 2\Delta^2_{ab}\mathscr{E}(\beta R_{ac}, \beta R_{bc}, \beta R_{c(ab)}) + [(ac), (bc)\} \\ & + \{-\Delta^2_{ab}\mathscr{D}(\beta R_{ac}, \beta R_{c(ab)}, \sphericalangle ac(ab)) \\ & - 2\Delta^2_{ab}\mathscr{F}(\beta R_{ab}, \beta R_{ac}, \beta R_{c(ab)}) \\ & + \Delta_{ab}\Delta_{ac}\Delta_{bc}\mathscr{G}(\beta R_{ab}, \beta R_{a(bc)}, \beta R_{(ab)(bc)}) \\ & + 2\Delta_{ab}\Delta_{ac}\Delta_{bc}\mathscr{H}(\beta R_{ab}, \beta R_{ac}, \beta R_{c(ab)}) \\ & + [(ba), (ac), (ca), (bc), (cb)]\}. \qquad (51)\end{aligned}$$

It is to be noted that the total number of permutations for the last four terms is six, compared with three for the first four terms of Eq. (51), since for the last terms permuting a and b or a and c or b and c is geometrically different for the arguments of the functions $\mathscr{D}$, $\mathscr{F}$, $\mathscr{G}$, and $\mathscr{H}$. In the first two terms on the right of Eq. (51) one recognizes again part of the two-body interaction for the pair ab, but the normalization constant has changed from $1 - \Delta^2_{ab}$ to $1 - \Delta^2_{abc}$. The difference between these terms of Eqs. (51) and (50) leads then to three-body effects which are purely due to overlap.

The final step in the evaluation of the relative second-order three-body interactions involves substituting the corresponding linear combinations of the basic integrals for the auxiliary functions in Eqs. (50) and (51) and computing the basic integrals for the arguments determined by the triangles considered. For the definitions of the basic integrals and further details on their evaluation see Jansen (1964).

We consider, as in the first-order calculations, triangles formed by a central atom and two of its twelve nearest neighbors in the hcp and fcc configurations. The relative second-order three-body interactions (41) are functions only of βR and of θ.

In Fig. 3, $\Delta E_2/E_2^{(0)}$ is given as a function of θ for solid argon ($\beta R = 2.4$). There appear to be two important contributions to $\Delta E_2/E_2^{(0)}$, due to:

(*i*) effects involving *diatomic exchange* only, i.e., terms arising from exchange of electrons between two of the three atoms on the triangle (curve I);

(*ii*) effects of *triatomic exchange*, i.e., those involving all three atoms on the triangle (curve II).

The total result for $\Delta E_2/E_2^{(0)}$, i.e., the sum of diatomic and triatomic exchange effects (curve I + II), has a surprising θ dependence: *it is practically identical with the one obtained for the first-order effect.* From the analytical form of the final result we conclude further that $\Delta E_2/E_2^{(0)}$ for neon, krypton, and xenon behaves in the same way as the first-order effects for these solids.

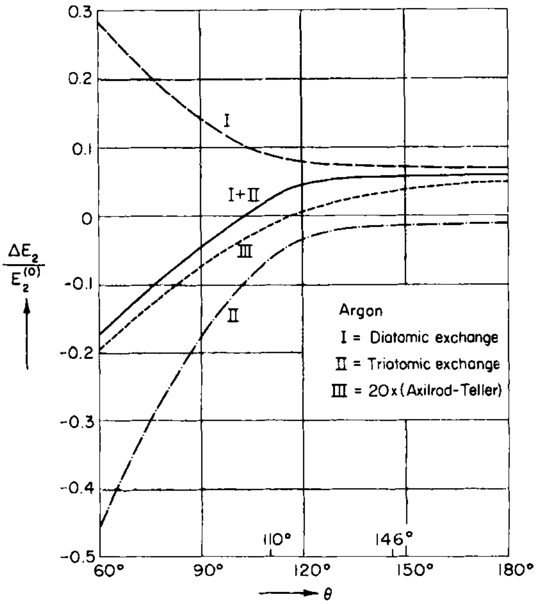


Fig. 3. Relative second-order three-body interaction $\Delta E_2/E_2^{(0)}$ for triangles $a=b=1$, c (units of nearest-neighbor distance) as a function of the angle θ between the sides a and b. The curves I and II denote diatomic and triatomic exchange contributions, respectively; (I + II) is the total second-order effect. Curve III represents the Axilrod–Teller (third-order) effect, magnified twenty times.

Since the total pair energy, $E_2^{(0)}$, for the triangles is always negative, we conclude that for values of θ between 60° and approximately 110° three-body second-order interactions *decrease* the interatomic attraction compared with an additive sum over pairs; for larger values of θ the three-body forces are *attractive. This change of sign agrees with the third-order Axilrod–Teller effect*, which is also plotted in Fig. 3 (curve III). It is observed that the second-order exchange forces are approximately twenty times larger than the third-order three-body interactions and that the exchange effect exhibits a much stronger θ dependence between $\theta = 90°$ and 120°.

The $\Delta E_1/E_1^{(0)}$ and $\Delta E_2/E_2^{(0)}$ curves are of the *same type*. Therefore, we can follow the same analysis as given in first order concerning stability. Since $E_1^{(0)}$ and $E_2^{(0)}$ have *opposite* sign, the conclusion is now reversed: *second-order three-body interactions favor the face-centered cubic configuration for rare gas crystals.* Summed over the triangles listed in Table I this difference, relative to the total two-body energy of the crystals, is of the order of a few per cent in favor of the fcc structure for argon, krypton, and xenon [on the basis of Eq. (39)].

The stability problem for rare gas crystals is herewith decided *in favor of the fcc configuration.* This is at once clear if we remember that the magnitude of the second-order crystal energy (attraction) is larger than that of first order (repulsion), in the ratio of approximately 2 : 1. Consequently, in spite of the approximate equality $\Delta E_1/E_1^{(0)} \approx \Delta E_2/E_2^{(0)}$, ΔE_2 weighs about twice as heavily as ΔE_1 for stability; i.e., the type of closest packing is determined by the three-atom interactions of *second* order. As we have seen, these favor the fcc configuration. It is remarkable to observe that the pair-equality $|E_2^{(0)}| \approx 2E_1^{(0)}$ for isosceles triangles consisting of a central atom and any two of its nearest neighbors also implies $|\Delta E_2| \approx 2|\Delta E_1|$ for the three-atom interactions.

We will investigate the properties of first- and second-order three-atom interactions in more detail in the following section. It is of interest to remark here that these results confirm the Rosen–Shostak (first-order) and Axilrod–Teller (third-order) results: (*i*) the relative three-atom interactions are negative for isosceles triangles with small opening θ; (*ii*) they are positive for isosceles triangles with large opening θ; (*iii*) the magnitude of the relative three-atom interactions at $\theta = 180°$ is smaller than at $\theta = 60°$.

C. Total Three-Atom Interactions and Rare Gas Crystal Stability

In Fig. 4 we compare the relative three-atom interactions in first and second orders for isosceles triangles $(\beta R, \theta)$ of argon atoms. Indicated also are the points of the curve referring to the nine hcp and the nine fcc triangles of Table I.

Three main conclusions can be drawn from the figure.

(*a*) $\Delta E_1/E_1^{(0)} \approx \Delta E_2/E_2^{(0)}$ *for all triangles considered.* This has the consequence that we may write the total (first-order plus second-order) relative three-body interactions for each triangle as follows:

$$\Delta E/E^{(0)} \equiv (\Delta E_1 + \Delta E_2)/(E_1^{(0)} + E_2^{(0)}) \approx \Delta E_1/E_1^{(0)} \approx \Delta E_2/E_2^{(0)}. \quad (52)$$

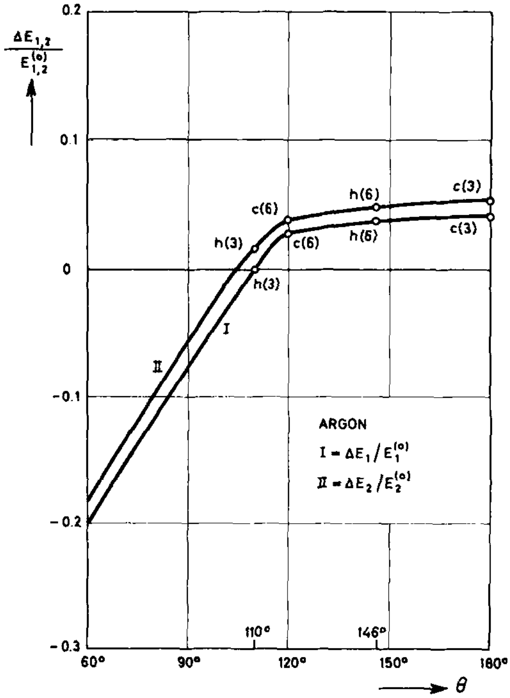


Fig. 4. Comparison between the relative first- and second-order three-body interactions for triangles $a = b = 1$, c (units of nearest-neighbor distance) of solid argon ($\beta R = 2.4$) as a function of the angle θ between the sides a and b. The nine different fcc and hcp triangles (Table I) are also indicated.

The following two possibilities are then to be distinguished.

(a1) $E^{(0)} = E_1^{(0)} + E_2^{(0)} < 0$, implying that the total pair-interactions between the atoms on the triangle are *attractive*. This situation applies for *van der Waals crystals*, such as those of the rare gases. Since $E_1^{(0)} > 0$, $E_2^{(0)} < 0$, we see that in this case ΔE has the same sign as ΔE_2, and we conclude:

The stability of rare gas crystals is determined by two-body interactions and by three-body interactions in second order of perturbation theory, i.e., the same order as the van der Waals forces themselves.

(a2) $E^{(0)} = E_1^{(0)} + E_2^{(0)} > 0$, implying that the total pair-interactions between the atoms on the triangle are *repulsive*. This situation applies for *alkali halide crystals*, where it is to be remembered that we have subtracted the electrostatic interactions between the ionic charges. The reason why

$E^{(0)}$ has now reversed its sign compared with van der Waals crystals is because the Madelung energy has *compressed* the crystal to the extent that the closest ions repel each other. In this case the total three-body energy ΔE has the same sign as ΔE_1, the first-order effect, and we conclude:

> *The stability of alkali halide crystals is determined by two-body interactions and by three-body interactions in first order of perturbation theory.*

(*b*) *First*-order three-body exchange interactions favor triangles with *small* opening θ.

(*c*) *Second*-order three-body exchange interactions favor triangles with *large* opening θ.

On the basis of these properties of three-body interactions the stability problem for rare gas solids can be solved. Since the pair energy, $E^{(0)}$, for each isosceles triangles is negative, ΔE has the same sign as ΔE_2; i.e., the total three-body effect *favors the face-centered cubic configuration for rare gas crystals.* To estimate the energy difference between the hcp and fcc structures relative to the pair energy of the crystal we proceed in the following way. Let ε denote the value of the pair energy at the equilibrium nearest-neighbor distance, R, in the lattice, and let Δ_θ denote the total (first-plus second-order) three-body energy for an isosceles triangle with opening θ. Then 3ε is the value of $E^{(0)}$ for $\theta = 60°$; for the other isosceles triangles it is sufficiently accurate to calculate $E^{(0)}$ on the basis of a Lennard-Jones (12, 6) potential. The values of $E^{(0)}$ at $\theta = 60°, 90°, 120°, 145°$, and $180°$ are then 3ε, $3\varepsilon\ (0.745)$, $3\varepsilon\ (0.691)$, $3\varepsilon\ (0.684)$, and $3\varepsilon\ (0.677)$, respectively.

To calculate Δ_θ we note that, for the isosceles triangles considered, $E_2^{(0)}$ is practically precisely $-2E_1^{(0)}$. This gives

$$\left(\frac{\Delta E_1 + \Delta E_2}{E_1^{(0)} + E_2^{(0)}}\right)E^{(0)} = \left(-\frac{\Delta E_1}{E_1^{(0)}} + 2\frac{\Delta E_2}{E_2^{(0)}}\right)E^{(0)}. \tag{53}$$

From Fig. 4 we determine values of Δ_θ of -0.48ε, -0.1341ε, ≈ 0, 0.1244ε, 0.1231ε, and 0.1219ε for $\theta = 60°, 90°, 109°, 120°, 145°$, and $180°$, respectively. By using lattice sums for the hexagonal and cubic structures we find that the total pair energy of these crystals is the same to within $10^{-2}\,\%$ and equal to $8.4\varepsilon N$, where N is the total number of atoms in the solid.

To determine the total three-body energy of the hcp and fcc structures we sum over all isosceles triangles with two atoms nearest neighbors of the third one. There are sixty-six such triangles formed by a central atom and any two of its twelve nearest neighbors. In the fcc structure there are twenty-four triangles at $\theta = 60°$, twelve at $90°$, twenty-four at $120°$, and six at $180°$.

In summing the three-body energy over the crystals it has to be noted that every equilateral triangle is counted three times in the summation, all other triangles only once. We obtain for the cohesive energy, E_{coh}, of the two structures:

$$(E_{\mathrm{coh}})_{\mathrm{fcc}}/N = 8.4\varepsilon + 8\Delta_{60} + 12\Delta_{90} + 24\Delta_{120} + 6\Delta_{180} \tag{54}$$

and

$$(E_{\mathrm{coh}})_{\mathrm{hcp}}/N = 8.4\varepsilon + 8\Delta_{60} + 12\Delta_{90} + 3\Delta_{109} + 18\Delta_{120} + 6\Delta_{145} + 3\Delta_{180}. \tag{55}$$

Hence, the relative difference is

$$\frac{(E_{\mathrm{coh}})_{\mathrm{fcc}} - (E_{\mathrm{coh}})_{\mathrm{hcp}}}{8.4\varepsilon N} = \frac{6\Delta_{120} + 3\Delta_{180} - (3\Delta_{109} + 6\Delta_{145})}{8.4\varepsilon}. \tag{56}$$

By substituting the values for Δ_θ, given earlier, we find

$$\frac{(E_{\mathrm{coh}})_{\mathrm{fcc}} - (E_{\mathrm{coh}})_{\mathrm{hcp}}}{8.4\varepsilon N} = 0.044. \tag{57}$$

Since $\varepsilon < 0$, this implies that the face-centered cubic lattice is more stable than the hexagonal close-packed configuration for solid argon by about 4% *of the pair cohesive energy*, thus largely exceeding the 0.01% difference in the pair energy which favors the hcp structure. If we take, in first approximation, only the difference between three fcc triangles at $\theta = 120°$ and three hcp triangles at $\theta = 109°$ [Eq. (39)], then we find the same value to within 0.1%. This shows again, as was noted earlier, that the *essential stabilizing factor* in going from the hcp to the fcc structure is the transition of three triangles from $\theta = 109°$ to $\theta = 120°$.

Besides the *difference* in cohesive energy between the two close-packed structures, it is also important to consider the effect of three-atom interactions on the cohesive energies of the two structures separately. We find directly, from Eqs. (54) and (55), that

$$(E_{\mathrm{coh}})_{\mathrm{fcc}} = 8.4N\varepsilon(1 - 0.206)$$

and

$$(E_{\mathrm{coh}})_{\mathrm{hcp}} = 8.4N\varepsilon(1 - 0.251),$$

implying that the cohesive energies for the cubic and hexagonal structures are *decreased*, in absolute value, by 21 and 25%, respectively, because of three-body interactions. The magnitude of this effect is surprising, since it is generally believed that the values of intermolecular potential parameters determined from gas data (PVT and viscosity measurements) are in good agreement with a pair-potential interpretation of the cohesive energies of rare gas solids (Dobbs and Jones, 1957; Hirschfelder *et al.*, 1954, Chapter 2). However, a recent analysis by Munn (1964) appears to indicate rather large deviations from pair-potential parameters for solid state properties.

In any case, the three-atom component of the cohesive energy for either structure is extremely sensitive to the precise value of $\Delta E/E^{(0)}$ for *small* values of θ, whereas the stability of the fcc structure is not at all sensitive in this region. This is, of course, due to the fact that there are no differences between the fcc and hcp structures for values of $\theta < 110°$ (each structure has eight triangles at $\theta = 60°$ and twelve triangles at $\theta = 90°$). If we would decrease $\Delta E_2/E_2^{(0)}$ at $\theta = 60°$ by 10% from -0.18 to -0.16 and at $\theta = 90°$ from -0.06 to -0.05, then we would obtain for the cohesive energies:

$$(E_{\mathrm{coh}})_{\mathrm{fcc}} = 8.4N\varepsilon(1 - 0.028);$$

and

$$(E_{\mathrm{coh}})_{\mathrm{hcp}} = 8.4N\varepsilon(1 - 0.073);$$

i.e., the three-atom components of the cohesive energies decrease by almost an order of magnitude. Consequently, no precise predictions are possible concerning these quantities, in view of the approximations of the effective-electron method.

Another quantity of interest in this connection is the energy of *vacancy formation* in solid argon. Foreman and Lidiard (1963) have recently compared data on the specific heat of solid argon for temperatures between 40°K and the melting point (84°K) with theoretical results on the basis of an anharmonic Einstein model for the ideal lattice. The difference between experimental and theoretical results was ascribed to the formation of vacancies in the crystal. These authors found a considerable discrepancy in the energy of vacancy formation with calculations based on two-body interactions (Nardelli and Repanai Chiarotti, 1960). The discrepancies can be explained quantitatively in terms of three-atom interactions (Jansen, 1963b, 1964), but this agreement may, in part, be fortuitous. The three-body component of the vacancy energy appears to be just *twice* that of the cohesive energy of solid argon; consequently, its value is very sensitive to the precise magnitude of $\Delta E/E^{(0)}$ at *small* values of the opening θ of isosceles triangles.

To conclude our considerations of rare gas crystal stability, we note again that the stability of the fcc structure depends primarily on the *type* of curve for $\Delta E/E^{(0)}$ as a function of the opening, θ, of isosceles triangles formed by a central atom and two of its nearest neighbors in the crystal.

This type agrees with the Axilrod–Teller relative third-order interaction except that it is more structure-sensitive, which is to be expected because of its exchange character. Further, as we mentioned before, if we take in the Rosen–Shostak results for three helium atoms $R_{ab} = R_{ac} = 1$ Å $= 1.98$ a.u., then $\Delta E_1/E_1^{(0)}$ is equal to -18% at $\theta = 60°$ and $+6.6\%$ at $\theta = 180°$, i.e., of the same order of magnitude as the Gaussian values for solid argon. This indicates that the specific choice of a Gaussian function for the charge distribution of an effective electron plays a "neutral" role in determining the θ dependence of the relative three-atom interactions.

Upon increasing the value of the dimensionless parameter βR, i.e., upon going to lighter rare gas atoms or expanding the crystal, the positive part of the $\Delta E/E^{(0)}$ curve approaches the horizontal axis. The lattice of solid helium (He^4) is relatively very wide because of large zero-point energy and the atom is relatively very small. In such a case three-atom interactions play no role for stability. The difference in lattice energy between the two close-packed configurations of rare gas crystals, although amounting to as much as 4%, is still small on an absolute scale, namely, of the order of 0.1 kcal/mole. It is therefore not excluded that the lighter solids may be observed in a metastable hexagonal packing by using special crystallization techniques (Barrett and Meyer, 1964; Meyer *et al.*, 1964).

We found that the rare gas crystals are constructed according to a minimum-energy principle involving pair-interactions and three-atom energies which follow simple symmetry principles. To illustrate these principles we consider again a central atom in the solid and its twelve nearest neighbors, of which six are on a hexagon in the central layer, three above and three below this layer. In the hcp structure the two triangles have the same orientation, whereas in the fcc configuration one is rotated by 60° with respect to the other. This rotation implies a very *small loss* in pair energy, but a *substantial gain* in three-atom interactions, involving the transition of three hcp triangles at $\theta = 109°28'$ to fcc triangles with slightly larger opening, namely 120°.

A large number of other van der Waals crystals are known to crystallize in close-packed structures. Some of them, for example nitrogen and carbon monoxide, exhibit a transition from fcc to hcp with increasing temperature. The same general type of three-atom exchange interactions must be expected to play a role for stability in these solids, supplemented by orientational two-body forces. It may be that we have an example in the crystal structure of methane, which is reported (Schallamach, 1939) to be fcc (i.e., the carbon atoms occupy the lattice points of a fcc structure). Very recent calculations by Kimel *et al.* (1964) based on atom-atom interactions yield, however, a *tetragonal* structure as the most stable one. The effect of orientational forces has been illustrated by Kihara (1963b), who built experimental models of molecules with large quadrupole moments and showed that these crystallize in the observed structures for sufficiently strong quadrupole interactions.

D. Alkali Halide Crystals. First- and Second-Order Three-Ion Exchange Interactions

We now turn to alkali halide crystals. We have assumed that there exists a close similarity between the stability problems of alkali halide and rare gas crystals in that the stability in both cases is determined by pair-interactions and by triplet interactions of exchange type between the ions or atoms, respectively.

However, the results for rare gas crystals cannot be taken over directly for alkali halide solids. First, the total three-ion interactions depend on βR, θ, the crystal structure, *and* on the additional parameter $\gamma = (\beta'/\beta)^2 > 1$, where β' is the Gaussian parameter for the *smaller* ion, β that for the *larger* ion, of one alkali halide. This additional parameter γ complicates the expressions for the three-body interactions considerably. Second, the Madelung energy for alkali halide crystals causes a compression of the lattice, so that the nearest-neighbor distances in the ionic solids are considerably smaller than those of rare gas crystals. As a consequence, we must consider the possible importance of *multiple-exchange* contributions to the three-ion interactions as well as contributions due to *next-nearest* and further neighbors of a central ion.

For the time being we will stay with the single-exchange approximation and consider corrections to this approximation later on. The following results have recently been published elsewhere (Lombardi and Jansen, 1964a,b).

1. *First-Order Calculations*

We follow the same procedure, with the same notation, as for three rare gas atoms (Section III,B). Let E_1 denote the *total* first-order energy for the (arbitrary) triplet *abc* of ions, where the expectation value is again determined with the zero-order wave function (34), and let $E_1^{(0)} = \langle H'_{abc}\rangle^{(0)}$ denote the sum of first-order interactions (repulsions) between the three *isolated* pairs of ions which constitute the triangle. We determine the relative three-ion component (36) of the first-order energy, i.e.,

$$\Delta E_1/E_1^{(0)} = (E_1 - E_1^{(0)})/E_1^{(0)},$$

as a function of the Gaussian parameter β for the larger ion (if a, b, and c represent two kinds of ion), of $\gamma = (\beta'/\beta)^2$, and of the triangular dimensions. The result for $\langle H'_{ab}\rangle/e^2$, equivalent to Eq. (37), is now

$$\begin{aligned}\langle H'_{ab}\rangle/e^2 = \frac{1}{R_{ab}} - \frac{1}{1-\Delta_{abc}^2}\{&-(1-\Delta_{bc}^2)G_{aa(b)} - (1-\Delta_{ac}^2)G_{bb(a)}\\ &+ (\Delta_{ab} - \Delta_{ac}\Delta_{bc})[G_{ab(a)} + G_{ab(b)}] + (\Delta_{ac} - \Delta_{ab}\Delta_{bc})G_{ac(b)}\\ &+ (\Delta_{bc} - \Delta_{ab}\Delta_{ac})G_{bc(a)} + (A_{abab} - A_{aabb}) + \Delta_{ac}(A_{abbc} - A_{abcb})\\ &+ \Delta_{bc}(A_{aabc} - A_{abac})\},\end{aligned} \tag{58}$$

where the integrals G and A are defined in the same way as for rare gas atoms. Because of the occurrence of the parameter γ, these integrals are considerably more complicated than for the special case $\gamma = 1$; for their specific values in case of Gaussian distributions of charge we refer to Lombardi and Jansen (1964a). Methods for evaluating these integrals have been developed by Shavitt (1963) and by Zimering (1965). It is easily verified that by putting $\gamma = 1$ the integrals are all reduced to the same form as those for rare gas atoms.

The first-order energy between a and b, *without* c present, $\langle H'_{ab}\rangle^{(0)}$, is obtained from Eq. (58) by removing ion c to infinity. The result is

$$\begin{aligned}\langle H'_{ab}\rangle^{(0)}/e^2 = \frac{1}{R_{ab}} + \frac{2}{1-\Delta_{ab}^2}\{&-G_{aa(b)} - G_{bb(a)} + \Delta_{ab}[G_{ab(a)} + G_{ab(b)}]\\ &+ (A_{abab} - A_{aabb})\}.\end{aligned} \tag{59}$$

The corresponding expressions for $\langle H'_{ac}\rangle^{(0)}$ and $\langle H'_{bc}\rangle^{(0)}$ are obtained from Eq. (59) by obvious substitutions. The *three-body* component of $\langle H'_{ab}\rangle/e^2$ is obtained by subtracting Eq. (59) from Eq. (58). Finally, we form the relative three-ion first-order energy, $\Delta E_1/E_1^{(0)}$, for each triplet (abc) of ions.

The total three-ion energies for the cesium chloride and sodium chloride configurations are then obtained by summing over all triplets in the two structures. In either lattice a central cation (anion) is surrounded by a first shell of anions (cations), followed by a second shell of cations (anions). Since the three-ion interactions are of short range, the main contributions arise from triangles of smallest dimensions. We indicate these triangles by the triplet of Gaussian parameters for the ions, where the first parameter represents the central ion. Thus, the smallest triangles are denoted by $(\beta\beta'\beta')$ and $(\beta'\beta\beta)$, followed by $(\beta\beta\beta)$ and $(\beta'\beta'\beta')$. The first two types refer to isosceles triangles formed by a central ion and two of its nearest neighbors, whereas the last two types denote isosceles triangles formed by a central ion and two of its *next*-nearest neighbors. We note that the contributions from the triangles $(\beta\beta\beta)$ and $(\beta'\beta'\beta')$ to the three-ion energy can be evaluated directly as for rare gas atoms.

For the stability problem we are essentially interested in those alkali halides which exhibit the cesium chloride structure under normal or moderate pressures, i.e., in the cesium and rubidium halides except the fluorides. For these halides βR lies in the neighborhood of 1.8 and γ assumes values between 1 and 2. We will find that in these cases the three-body component of the lattice energy is determined essentially by contributions from the triangles of types $(\beta'\beta\beta)$ and $(\beta\beta'\beta')$, whereas the triplets $(\beta\beta\beta)$ and $(\beta'\beta'\beta')$ are relatively unimportant. On the other hand, the more dissimilar the ions are in size, i.e., the larger the value of γ, the more important contributions from triplets $(\beta\beta\beta)$ of the larger ions to the three-body energy become.

In Figs. 5 and 6 we give the results for the relative first-order three-body energy $\Delta E_1/E_1^{(0)}$ as a function of the opening, θ, of the isosceles triangles $(\beta'\beta\beta)$ and $(\beta\beta'\beta')$ for $\beta R = 1.8$ and $\gamma = 1$, 1.25, 1.50, 1.75, and 2.00. The results were obtained, from Eqs. (59) and (58), with an IBM-1620 computer. We draw the following conclusions.

(1) $\Delta E_1/E_1^{(0)}$ as a function of θ *exhibits the same general behavior as for rare gas crystals*: the relative three-body energy is *negative* for triangles with small opening, increases rapidly with θ until $\theta \sim 120°$, then flattens off very markedly, assuming positive values for triangles with large opening.

(2) For triangles $(\beta'\beta\beta)$, i.e., those with the *smaller* ion at the center, increasing γ has a negligible effect for small θ, whereas the three-body energy is quenched considerably for large values of θ.

(3) The *opposite* behavior from (2) is exhibited by triangles $(\beta\beta'\beta')$, i.e., those with the *larger* ion at the center.

Since $E_1^{(0)} > 0$, we conclude that in all cases first-order three-body interactions *favor triangles with small opening* θ, as in the case of rare gas crystals.

If, on the other hand, the cation is much smaller than the anion, i.e., if $\gamma \gg 1$, then we expect more drastic changes in the behavior of $\Delta E_1/E_1^{(0)}$. This is shown in Figs. 7 and 8 where $\Delta E_1/E_1^{(0)}$ is given for triplets $(\beta'\beta\beta)$ and $(\beta\beta'\beta')$, as a function of θ, for $\beta R = 1.5$ and $\gamma = 2$, 4, 10, 20, and an extreme value of 100. For comparison, the values for $\gamma = 1$ are also reported.

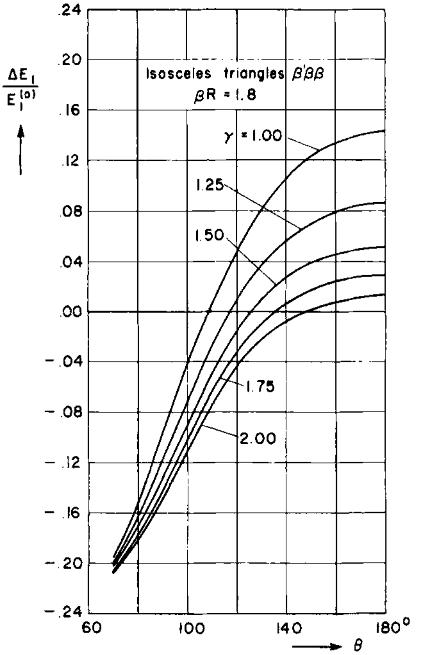


Fig. 5. Relative first-order three-ion interactions $\Delta E_1/E_1^{(0)}$ for isosceles triangles $(\beta'\beta\beta)$, i.e., with the smaller ion at the center, as a function of the opening θ at the central ion, for $\beta R = 1.8$ and $\gamma = (\beta'/\beta)^2 = 1.00$, 1.25, 1.50, 1.75, and 2.00.

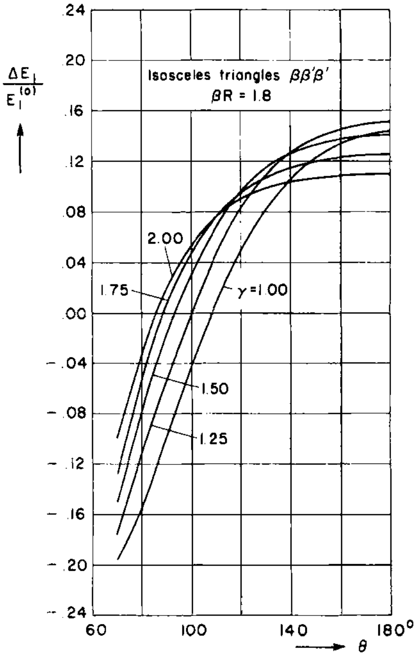


Fig. 6. Relative first-order three-ion interactions $\Delta E_1/E_1^{(0)}$ for isosceles triangles $(\beta\beta'\beta')$, i.e., with the larger ion at the center, as a function of the opening θ at the central ion, for $\beta R = 1.8$ and $\gamma = 1.00$, 1.25, 1.50, 1.75, and 2.00.

From Figs. 7 and 8 we draw the following conclusions.

(1*a*) In the case of triplets $(\beta'\beta\beta)$ we see (Fig. 7) that $\Delta E_1/E_1^{(0)}$ becomes practically independent of γ for $\theta > 150°$, whereas the values for small θ increase rapidly with γ.

(2*a*) In the case of triplets $(\beta\beta'\beta')$, with a much larger ion at the center, we see (Fig. 8) that the three-body interactions are independent of θ for $\gamma \geq 10$. For $\gamma \sim 4$ the values increase very sharply for the smallest openings and then remain practically constant.

We see that with increasing γ the three-body energy is *quenched* considerably and *loses its sensitivity* with respect to the crystal structure. From this we anticipate that in the case of very dissimilar ions the main three-body contribution will arise from triangles $(\beta\beta\beta)$, i.e., those formed by a large ion at the center and two of its *next*-nearest neighbors.

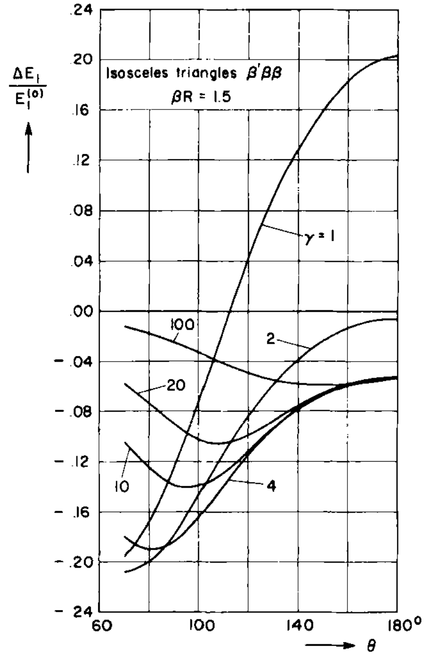


Fig. 7. Relative first-order three-ion interactions $\Delta E_1/E_1^{(0)}$ for isosceles triangles ($\beta'\beta\beta$), i.e., with the smaller ion at the center, as a function of the opening θ at the central ion, for $\beta R = 1.5$ and $\gamma = (\beta'/\beta)^2 = 1, 2, 4, 10, 20$, and 100.

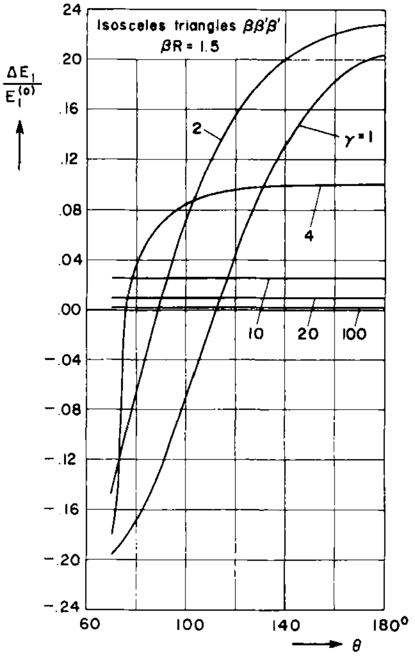


Fig. 8. Relative first-order three-ion interactions $\Delta E_1/E_1^{(0)}$ for isosceles triangles ($\beta\beta'\beta'$), i.e., with the larger ion at the center, as a function of the opening θ at the central ion, for $\beta R = 1.5$ and $\gamma = 1$, 2, 4, 10, 20, and 100.

In addition to the types of isosceles triangles considered above, we will later also evaluate contributions to the three-body energy from *non*isosceles triangles in the sodium chloride and cesium chloride configurations.

2. *Second-Order Calculations*

For the second-order energy of a triplet (*abc*) of ions we must evaluate the same expression (40) as for rare gas atoms. The quantity of direct interest is again the relative three-body second-order energy $\Delta E_2/E_2^{(0)}$ given by Eq. (41).

We have found that for rare gas atoms the evaluation of $\Delta E_2/E_2^{(0)}$ constitutes a problem of considerable complexity. In the case of alkali halide ions this task is even much more formidable in view of the occurrence of the parameter γ in the equations. We remember, however, that we are here essentially interested in heavy alkali halides, and for this limiting case the second-order three-ion interactions can be evaluated without detailed calculations. We found for first-order interactions of such solids, characterized by γ values between 1 and 2 and by $\beta R \sim 1.8$, that the results are very similar to those for rare gas atoms. It must be expected that this similarity *extends* to second-order interactions. Specifically, we assume that in this range of values for γ and βR the relation $\Delta E_1/E_1^{(0)} \approx \Delta E_2/E_2^{(0)}$ *holds also for three-ion interactions.* Further, for values of $\gamma \gg 1$ we have found that the three-ion first-order interactions are strongly quenched; the three-ion component of the crystal energy is then mainly due to triplets $(\beta\beta\beta)$ and the relation $\Delta E_1/E_1^{(0)} \approx \Delta E_2/E_2^{(0)}$ is again valid. By interpolation the relative first- and second-order three-ion interactions will be taken as equal over the whole range of values for βR and γ of alkali halide crystals. Since the second-order pair-energy $E_2^{(0)}$ for any triplet is negative, the symmetry properties of second-order three-ion interactions are just the *reverse* of those in first order in that triangles with *large* opening are now favored.

On the basis of these approximations we can now complete the stability analysis for the alkali halide crystals. Before doing so, however, we must investigate possible corrections to the single-exchange effective-electron method because of *multiple-exchange* effects. In principle, such corrections apply for rare gas crystals as well, although in that case we expect them to be less important, because of the relatively large values for βR.

E. Alkali Halide and Rare Gas Crystals. Double-Exchange Contributions to Three-Body Energy

The results which we have so far obtained for the relative three-atom and three-ion interactions were based on a one-electron approximation for the atoms and ions. This implies that we have taken into account only contributions due to exchange of a *single* pair of electrons between the same pair of atoms (ions). As the distance between them decreases, it must be expected that effects due to multiple exchange (double exchange in the first place) become more important. Consequently, we must verify that the results of the previous sections do not change essentially if double exchange is taken into account. In particular, the symmetry properties of three-atom or three-ion interactions should be *stable* against double exchange.

To calculate double-exchange contributions it is sufficient to investigate the special case $\gamma = 1$. Consider a triplet (abc) of identical Gaussian ions or atoms with *two* effective electrons per ion (opposite spins), counterbalanced by nuclear charges of plus two. The zero-order wave function is (Slater determinant):

$$\Psi^{(0)} = [6!(1 - \Delta_{abc}^2)]^{-1/2} \det\{\varphi_a[1]\alpha_1\varphi_a[2]\beta_2\varphi_b[3]\alpha_3\varphi_b[4]\beta_4 \times \varphi_c[5]\alpha_5\varphi_c[6]\beta_6\}, \tag{60}$$

where 1–6 numbers the electrons and where α and β are spin functions (symbol β not to be confused with the Gaussian parameter). The atomic wave functions $\varphi(r)$ are again of the form shown in Eq. (33), and Δ_{abc}^2 is defined by Eq. (35). The perturbation Hamiltonian H'_{abc} is again a sum of perturbations between the three pairs of atoms, in this case with two electrons per atom.

The resulting expression for $\langle H'_{ab}\rangle/e^2$ is more complicated than the corresponding Eq. (37) for the single-exchange approximation; it can be written, noting that for identical atoms $G_{aa(b)} = G_{bb(a)}$, $G_{ab(a)} = G_{ab(b)}$, and $A_{aaab} = A_{bbba}$, as

$$\begin{aligned}\langle H'_{ab}\rangle/e^2 = \frac{4}{R_{ab}} + \frac{1}{(1 - \Delta_{abc}^2)^2} \{&-4G_{aa(b)}P + 8G_{ab(a)}Q + 4(1 + [ab])G_{ac(b)}R \\ &+ 2A_{abab}S + 2(1 + [ab])A_{acab}T + 2A_{abba}U + 2(1 + [ab])A_{aaab}V \\ &+ 2(1 + [ab])A_{acba}W + 2A_{abcc}X\},\end{aligned} \tag{61}$$

where the symbol $[ab]$ stands for the operation of permuting a and b, and where the symbols P to X are defined as follows:

$$\begin{aligned}P = 2 &- 2\Delta_{ab}^2 - 3\Delta_{ac}^2 - 3\Delta_{bc}^2 + 4\Delta_{ab}\Delta_{ac}\Delta_{bc} + \Delta_{ab}^2\Delta_{bc}^2 + \Delta_{ab}^2\Delta_{ac}^2 + 2\Delta_{ac}^2\Delta_{bc}^2 \\ &+ \Delta_{ac}^4 + \Delta_{bc}^4 - 2\Delta_{ab}\Delta_{ac}\Delta_{bc}^3 - 2\Delta_{ab}\Delta_{bc}\Delta_{ac}^3;\\ Q = \Delta_{ab} &- \Delta_{ac}\Delta_{bc} - \Delta_{ab}^3 - \Delta_{ab}\Delta_{bc}^2 - \Delta_{ab}\Delta_{ac}^2 + \Delta_{ac}\Delta_{bc}^3 + \Delta_{bc}\Delta_{ac}^3 + 3\Delta_{ab}^2\Delta_{ac}\Delta_{bc} \\ &- 2\Delta_{ab}\Delta_{ac}^2\Delta_{bc}^2;\\ R = \Delta_{ac} &- \Delta_{ab}\Delta_{bc} - \Delta_{ab}^2\Delta_{ac} - \Delta_{bc}^2\Delta_{ac} - \Delta_{ac}^3 + \Delta_{ab}\Delta_{bc}^3 + \Delta_{ab}^3\Delta_{bc} + 3\Delta_{ab}\Delta_{ac}^2\Delta_{bc} \\ &- 2\Delta_{ab}^2\Delta_{ac}\Delta_{bc}^2;\end{aligned}$$

$$S = 2 - 2\Delta_{ac}^2 - 2\Delta_{bc}^2 - \Delta_{ab}^2 + 2\Delta_{ab}\Delta_{ac}\Delta_{bc} + \Delta_{ac}^2\Delta_{bc}^2;$$

$$T = -2\Delta_{bc} + \Delta_{ab}\Delta_{ac} + 2\Delta_{bc}^3 + \Delta_{ab}^2\Delta_{bc} + \Delta_{ac}^2\Delta_{bc} - 3\Delta_{ab}\Delta_{ac}\Delta_{bc}^2;$$

$$U = -1 + \Delta_{ac}^2 + \Delta_{bc}^2 + 2\Delta_{ab}^2 - 4\Delta_{ab}\Delta_{ac}\Delta_{bc} + \Delta_{ac}^2\Delta_{bc}^2;$$

$$V = -\Delta_{ab} + \Delta_{ac}\Delta_{bc} - \Delta_{ac}\Delta_{bc}^3 + \Delta_{ab}\Delta_{bc}^2;$$

$$W = \Delta_{ac} + \Delta_{ab}\Delta_{ac} - 2\Delta_{ab}^2\Delta_{ac} - 2\Delta_{ac}\Delta_{bc}^2 - \Delta_{ac}^3 + 3\Delta_{ab}\Delta_{bc}\Delta_{ac}^2;$$

and

$$X = \Delta_{ac}\Delta_{bc} - \Delta_{ab}\Delta_{bc}^2 - \Delta_{ab}\Delta_{ac}^2 + \Delta_{ab}^2\Delta_{ac}\Delta_{bc}. \tag{62}$$

The equations for $\langle H'_{ac}\rangle/e^2$ and $\langle H'_{bc}\rangle/e^2$ are obtained from Eqs. (61) and (62) by permutations.

The first-order energy for the pair (ab), without c present, is then given by Eq. (61) with atom c removed to infinity. The result, corresponding to Eq. (38) is:

$$\langle H'_{ab}\rangle^{(0)}/e^2 = \frac{4}{R_{ab}} - \frac{4}{(1-\Delta_{ab}^2)}\{2G_{aa(b)} - 2\Delta_{ab}G_{ab(a)} - (A_{abab} - A_{aabb})\}$$
$$+ \frac{2}{(1-\Delta_{ab}^2)^2}[\Delta_{ab}^2 A_{abab} + A_{aabb} + 2\Delta_{ab}A_{aaab}]. \tag{63}$$

The interactions between a and c, or those between b and c, follow from Eq. (63) by permutations.

In Fig. 9 the results for $\Delta E_1/E_1^{(0)}$ are given for $\beta R = 1.0$, 1.6, and 2.0, as a function of the opening θ of isosceles triangles, both for one and for two effective electrons.

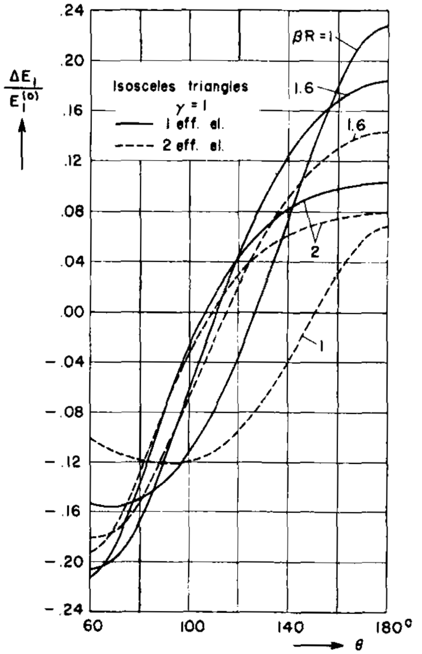


Fig. 9. Relative first-order three-body interactions $\Delta E_1/E_1^{(0)}$, in the one and two effective-electron approximations, for isosceles triangles, as a function of their opening θ, for $\gamma = 1$ and $\beta R = 1$, 1.6, and 2.0. The drawn-out curves include double-exchange contributions to the three-body interaction.

We draw the following conclusions.

(*a*) *No qualitative changes occur for* $\Delta E_1/E_1^{(0)}$, as a function of θ, if double exchange is taken into account. In particular, the symmetry properties of the three-body interactions remain *the same* even for values as small as $\beta R = 1$.

(*b*) Double exchange *quenches* $\Delta E_1/E_1^{(0)}$ both at small and at large openings of the triangles, but the quenching effect decreases rapidly with increasing βR. These results confirm validity of the single-exchange approximation for rare gas crystals ($\beta R \geq 2$).

(*c*) The double-exchange analysis supports the assumed similarity between stability of rare gas and alkali halide crystals.

In view of these results it is justified to carry out the stability analysis on the basis of a *single-exchange* approximation.

F. The Stability of Alkali Halide Crystals

The results of the previous sections will now be applied to the stability of alkali halide crystals. [We will closely follow the text of Lombardi and Jansen (1964a).] As a simplification, we limit ourselves to the absolute zero of temperature and neglect the effect of zero-point energy on stability, so that we only have to compare the *static* lattice energies of the cesium chloride and the sodium chloride configurations for the different alkali halides.

Consider a crystal of N cations and N anions; the static lattice energy, $E(r_1, r_2, \ldots, r_{2N})$, is defined as the difference between the energy of the crystal, for a fixed configuration $r_1, r_2, \ldots, r_{2N}$ of nuclei, and the energy of the $2N$ constituents at infinite separations and at absolute zero temperature. We write the lattice energy formally as a series of terms depending upon the position coordinates of increasing numbers of ions as follows:

$$E(r_1, r_2, \ldots, r_{2N}) = E\{2\} + E\{3\} + E\{4\} + \cdots, \tag{64}$$

where $E\{2\}$ contains all terms of E which depend on the coordinates of only two ions; in $E\{3\}$ all terms are collected which are simultaneously functions of the coordinates of three ions, etc. We assume that this expansion may be terminated with the three-particle function $E\{3\}$ for molecular crystals and ionic solids.

Expressions for the components $E\{2\}$, $E\{3\}$, etc., of E in terms of the interactions between the ions can easily be given. Evidently, $E\{2\}$ is just the sum of interactions for the *isolated pairs* of ions in the crystal. Further,

$E\{3\}$ is the limiting value of $E - E\{2\}$ if all simultaneous interactions between more than three ions are discarded, i.e.,

$$E\{3\} = \sum_{a<b<c} [E(abc) - \{E^{(0)}(ab) + E^{(0)}(ac) + E^{(0)}(bc)\}]$$
$$\equiv \sum_{a<b<c} \Delta E(abc), \tag{65}$$

where $E(abc)$ denotes the total interaction energy for the *isolated* triplet (abc) and $E^{(0)}(ab)$ the interaction for the isolated pair (ab); the summations extend over all possible triplets. The pair interactions $E^{(0)}(ab)$ and the triplet interactions $\Delta E(abc)$ are evaluated in first and second orders of perturbation theory. We write $E^{(0)} = E_1^{(0)} + E_2^{(0)}$ and $\Delta E = \Delta E_1 + \Delta E_2$ for any pair or triplet of ions; the subscripts 1 and 2 denote the orders of perturbation, as in the previous sections.

We first consider the *pair energy* $E\{2\}$ for the crystal. Since accurate theoretical expressions for the pair-interactions between ions are not available, we use the empirical information that the *total pair energy* of alkali halide crystals may be calculated with a model of electrostatic interactions between *point* charges for the ions, supplemented by first-order repulsions between the closed shells. We remark that the van der Waals (second-order) interactions between the ions are only indirectly represented in this model. Since we have ascribed the structure dependence of the repulsive parameters, analyzed by Tosi and Fumi (1962), to the effect of *three-body* interactions, we write, accordingly, the total pair energy $E\{2\}$ for each alkali halide and either crystal structure as a sum of Madelung energy, M, and of the total first-order repulsion between the ions, i.e., as

$$E\{2\} = M + \sum_{a<b} E_1^{(0)}(ab). \tag{66}$$

Further, let $\Delta E = \Delta E_1 + \Delta E_2$ represent the three-body energy for an arbitrary triplet and $\Delta E_1/E_1^{(0)}$ and $\Delta E_2/E_2^{(0)}$ the relative first- and second-order three-body interactions. With the assumption $\Delta E_1/E_1^{(0)} \approx \Delta E_2/E_2^{(0)}$ we can write

$$\Delta E = (\Delta E_1/E_1^{(0)})E_1^{(0)} + (\Delta E_2/E_2^{(0)})E_2^{(0)} \approx (\Delta E_1/E_1^{(0)})(E_1^{(0)} + E_2^{(0)}), \tag{67}$$

where $E_1^{(0)} + E_2^{(0)} = E^{(0)}$ is the total (first- plus second-order) pair energy of the triplet. It is important to note, from Eq. (67), that in this case the van der Waals pair-interactions $E_2^{(0)}$ must explicitly be taken into account. The three-body crystal energy $E\{3\}$ is the sum of Eq. (67) over all possible triplets of ions.

To evaluate $E\{3\}$ for the cesium chloride and the sodium chloride structures, we first classify the different triangles formed by a central ion and two neighbors of the first few shells in the two configurations. Before, we have indicated such triangles by the corresponding triplet of Gaussian parameters, but for what follows a more specific notation is needed. We denote a cation by C and an anion by A, and add subscripts 0, 1, and 2 to distinguish between different shells around the central ion. The subscript zero refers to the central cation or anion.

In Tables V and VI the numbers, per ion, and the types of different triangular configurations are given for the sodium chloride and cesium chloride structures. The sides R_{ab}, R_{ac}, and R_{bc} are expressed in units of nearest-neighbor distance R; θ is the angle between R_{ab} and R_{ac}. The type of triangle $A_0C_1A_2$, for example, denotes a triangle formed by a central anion, a cation of the first shell, and an anion belonging to the second shell of neighbors.

TABLE V

Numbers and Types of Triangles, *per Ion*, in the Sodium Chloride Configuration[a]

Type	Number per ion	R^2_{ab}	R^2_{ac}	R^2_{bc}	$\cos^2\theta$	θ
$A_0C_1C_1$	12	1	1	2	0	90°
$(C_0A_1A_1)$	3	1	1	4	1	180°
$A_0A_2A_2$	8	2	2	2	1/4	60°
$(C_0C_2C_2)$	12	2	2	4	0	90°
	24	2	2	6	1/4	120°
	6	2	2	8	1	180°
$A_0C_1A_2$	24	1	2	3	0	90°
$(C_0A_1C_2)$	24	1	2	5	1/2	135°
$A_0C_1C_3$ $(C_0A_1A_3)$	24	1	3	6	1/3	125° 15′
$A_0A_2C_3$	48	2	3	5	0	90°
$(C_0C_2A_3)$	24	2	3	9	2/3	144° 44′

[a] The sides of the triangles are in units of nearest-neighbor distance; θ is the angle between R_{ab} and R_{ac}.

As is seen from these tables, the first two types of triplets in the sodium chloride structure and the first three types of triplets in the cesium chloride structure form isosceles triangles. The remaining types of triangles have three different sides; we will first verify that the contributions to the three-body energy due to these nonisosceles triangles are small, so that we then may restrict ourselves to isosceles triangles only.

TABLE VI

NUMBERS AND TYPES OF TRIANGLES, *per Ion*, IN THE CESIUM CHLORIDE CONFIGURATION[a]

Type	Number per ion	R^2_{ab}	R^2_{ac}	R^2_{bc}	$\cos^2\theta$	θ
$A_0C_1C_1$	12	1	1	4/3	1/9	70° 32′
$(C_0A_1A_1)$	12	1	1	8/3	1/9	109° 28′
	4	1	1	4	1	180°
$A_0A_2A_2$	12	4/3	4/3	8/3	0	90°
$(C_0C_2C_2)$	3	4/3	4/3	16/3	1	180°
$A_0A_3A_3$	8	8/3	8/3	8/3	1/4	60°
$(C_0C_3C_3)$	12	8/3	8/3	16/3	0	90°
	24	8/3	8/3	8	1/4	120°
	6	8/3	8/3	32/3	1	180°
$A_0A_2A_3$	24	4/3	8/3	4	0	90°
$(C_0C_2C_3)$	24	4/3	8/3	20/3	1/2	135°
$A_0C_1A_2$ $(C_0A_1C_2)$	24	1	4/3	11/3	1/3	125° 15′
$A_0C_1A_3$	48	1	8/3	11/3	0	90°
$(C_0A_1C_3)$	24	1	8/3	19/3	2/3	144° 44′

[a] The sides of the triangles are in units of nearest-neighbor distance; θ is the angle between R_{ab} and R_{ac}.

We have determined the relative first-order three-body energy $\Delta E_1/E_1^{(0)}$, as a function of βR, for the congruent types of triangles $A_0C_1A_2$ (sodium chloride) and $A_0A_2A_3$ (cesium chloride), as well as for the types $A_0C_1A_2$ and $A_0C_1A_3$ (cesium chloride). For simplicity, we considered only the case $\gamma = 1$. The results are given in Fig. 10. For comparison, curve *a* of Fig. 10 refers to an *isosceles* triangle with $\theta = 90°$.

We draw the following conclusions.

(*a*) $\Delta E_1/E_1^{(0)}$ is quenched for $\theta = 90°$ as the third ion is moved away from the other two; this is apparent from curves *a*, *b*, and *c*. Also, $\Delta E_1/E_1^{(0)}$ decreases rapidly with increasing βR.

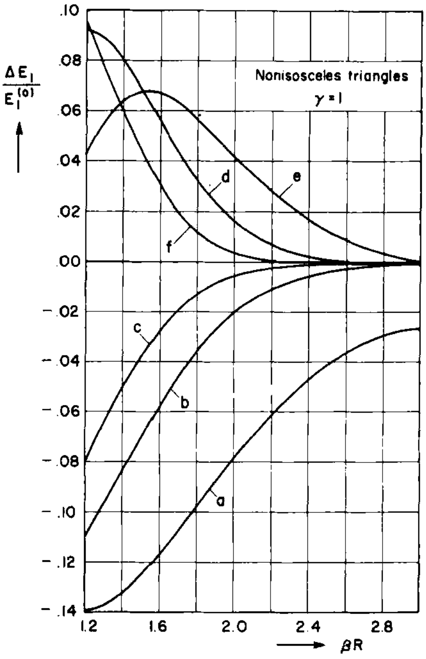


Fig. 10. Contributions of nonisosceles triangles of ions in the NaCl and CsCl configurations, for the special case $\gamma = 1$, to the relative first-order three-ion energy $\Delta E_1/E_1^{(0)}$, as a function of βR. Curve *a* refers to isosceles rectangular triangles for comparison, and curves *b* and *c* to rectangular triangles with ratios 1 : 2 : 3 and 1 : 8/3 : 11/3, respectively, for the squares of their sides. For curves *d*, *e*, and *f* these ratios are, respectively, 1 : 2 : 5, 1 : 4/5 : 11/3, and 1 : 8/3 : 19/3.

(*b*) Contributions from nonisosceles triangles in either structure *tend to cancel each other*. For example, consider the type $A_0C_1A_2$ (sodium chloride); there are twenty-four such triangles with $\theta = 90°$ (curve *b*) and twenty-four with $\theta = 135°$ (curve *d*). These two contributions are practically equal and of opposite sign over the whole range of values of βR. In the same way, considering the types $A_0C_1A_2$ and $A_0C_1A_3$ (cesium chloride), we note that there are twenty-four triangles with $\theta = 125° 15'$ (curve *e*), and twenty-four with $\theta = 144° 44'$ (curve *f*), which practically cancel against forty-eight triangles with $\theta = 90°$ (curve *c*). The same considerations apply to the types $A_0C_1C_3$ and $A_0A_2C_3$ (sodium chloride) and to the type $A_0A_2A_3$ (cesium chloride).

It is thus seen that the contributions from nonisosceles triangles are *small, of the same order in the two structures, and, moreover, that they tend to cancel each other in either structure because of symmetry properties of three-body interactions.*

We will now undertake the stability analysis, restricting ourselves to the *isosceles* triangles of Table V (sodium chloride structure) and of Table VI (cesium chloride structure). A number of qualitative features of the differences between the two structures with respect to stability can already be obtained by comparing triangles of Table V and VI and applying symmetry properties of three-body interactions. However, we will postpone a more general discussion until after the complete numerical results have been presented.

The stability analysis is carried out in the following four steps.

(*a*) *For each alkali halide and for both structures we determine the total three-body energy* (65) *for* N *cations and* N *anions.* The values of γ and βR are taken from Table III. To determine ΔE for each triangle from Eq. (67), we must know its *total* pair energy, including contributions from van der Waals interactions. For rare gas crystals a pair-potential function is taken as basis, e.g., a Lennard-Jones (12, 6) potential, placing nearest neighbors at a distance where the potential has its minimum. For alkali halides such pair-potentials [cf. Varshni and Shukla (1961) for detailed references] are not known with precision.

We estimate that, on the basis of a Lennard-Jones (12, 6) potential, the nearest-neighbor distance may be taken between 0.85σ and 0.90σ, where σ is the distance for zero potential. The ratio between repulsive and attractive interactions for such a pair varies between -2.7 and -1.9. This would imply that $E_1^{(0)} + E_2^{(0)}$ for a pair of nearest neighbors varies between $0.63E_1^{(0)}$ and $0.47E_1^{(0)}$. We write for the *total pair energy* of the triplet $E_1^{(0)}(1 - b)$, with b between 0.4 and 0.6, approximately.

On the other hand, if next-nearest neighbors are at distances comparable to the nearest-neighbor separation in rare gas crystals, then the relation $E_2^{(0)} \approx -2E_1^{(0)}$ holds, so that $E_1^{(0)} + E_2^{(0)} \approx -E_1^{(0)}$ in this case. Contributions from triangles at much larger distances will be neglected.

On the basis of these estimates we can evaluate ΔE for each triplet in terms of its total first-order pair energy $E_1^{(0)}$. To sum over the different triangles we must relate their first-order energies. Let α_1 denote the pair repulsion between nearest neighbors, α_2 the repulsion between next-nearest neighbors, and α_3 that between third neighbors in the crystal. We equate α_2/α_1 and α_3/α_1 to the corresponding ratios between first-order interactions in the Gaussian model. Since $E_1^{(0)}$ for any triplet considered is a function of α_1, α_2, and α_3, we can express in this way $E_1^{(0)}$ for each triplet as a function of α_1 only.

(*b*) *The total first-order pair energy* E{2} *is then evaluated*, from Eq. (66), by adding to the Madelung energy the sum of all pair repulsions as a function of α_1, using Gaussian ratios for distant neighbors.

(*c*) From (*a*) and (*b*) we obtain the *sum of all pair and triplet interactions*, as a function of α_1, for the two structures. This sum, for the stable structure, is put equal to the experimental value of the lattice energy as determined by the Born–Haber cycle. We use this equality to determine the nearest-neighbor repulsion α_1.

(*d*) With the value of α_1 obtained from (*c*), we calculate the *lattice energy of the other structure*. The energy difference between the two structures is then used to evaluate also transition pressures from the sodium chloride to the cesium chloride structure.

The results of steps (*a*) and (*b*) for all alkali halides in the sodium chloride and cesium chloride configurations are reported in Tables VII and VIII, respectively. The second column of each table gives the total pair repulsion, in units of the nearest-neighbor repulsion α_1. In the third column we list the total three-body contributions from $A_0C_1C_1$ and $C_0A_1A_1$ triangles, whereas the fourth column gives the total three-body interactions for triangles $C_0C_2C_2$ and $A_0A_2A_2$ in either structure. Finally, in the fifth column of Table VIII we list the contributions from triangles $A_0A_3A_3$ and $C_0C_3C_3$ in the cesium chloride configuration.

From Tables VII and VIII we draw the following main conclusions

(1) The three-body interactions for triplets involving *nearest neighbors* ($A_0C_1C_1$ and $C_0A_1A_1$) increase the *attractive* forces in either structure. The contribution is larger for the *cesium chloride* configuration, *which is thus favored by these types of triangles.*

(2) The three-body interactions for triplets involving *next-nearest neighbors* ($A_0A_2A_2$ and $C_0C_2C_2$) in the sodium chloride structure and next-nearest as well as third neighbors in the cesium chloride structure ($A_0A_2A_2$, $C_0C_2C_2$, $A_0A_3A_3$, and $C_0C_3C_3$) increase the *repulsive* forces in either structure. The contribution is smaller (less repulsive) for the *sodium chloride* configuration, *which is thus favored by these types of triangles.*

(3) When the ions become more dissimilar in size, i.e., when γ increases, the three-body interactions involving *next-nearest neighbors* become increasingly more important with respect to those involving first neighbors.

TABLE VII

TOTAL PAIR REPULSION (COLUMN 2) AND TOTAL THREE-BODY ENERGY INVOLVING NEAREST NEIGHBORS (COLUMN 3) AND NEXT-NEAREST NEIGHBORS (COLUMN 4) FOR THE ALKALI HALIDES IN THE SODIUM CHLORIDE CONFIGURATION[a]

Alkali halide	Sodium chloride configuration		
	Pair repulsion	Three-body energy	
		$A_0C_1C_1$; $C_0A_1A_1$	$A_0A_2A_2$; $C_0C_2C_2$
LiF	11.68	−4.46	4.36
NaF	6.30	−1.91	0.24
KF	6.82	−2.41	0.64
RbF	8.09	−3.07	1.48
CsF	10.21	−3.88	2.64
LiCl	37.23	−7.37	18.46
NaCl	11.06	−4.44	3.32
KCl	7.20	−2.81	0.83
RbCl	6.96	−2.92	0.72
CsCl	7.03	−2.96	0.53
LiBr	55.33	−8.10	27.47
NaBr	14.56	−5.40	5.06
KBr	8.31	−3.23	1.47
RbBr	7.55	−3.18	1.04
CsBr	7.16	−3.13	0.83
LiI	75.50	−8.48	37.59
NaI	18.49	−6.21	7.13
KI	9.65	−3.64	2.20
RbI	8.30	−3.37	1.39
CsI	7.47	−3.22	1.00

[a] All results are expressed in units of nearest-neighbor repulsion α_1 except those in column 2 where the unit is $\alpha_1(1\text{-}b)$. The parameter b stands for the ratio between first- and second-order pair interactions for a triplet of nearest neighbors; $b \approx 0.5$ approximately.

(4) In addition to the difference in Madelung energy between the two structures, *the stability is determined by a balance between pair repulsions, three-body attractive forces, and three-body repulsive forces*; these different components of the crystal energy depend on the values of βR and γ. For large values of γ we predict the sodium

chloride structure to be the stable one. As will be shown, this influence of γ explains the difference in crystal structure between cesium fluoride ($\gamma \approx 3.5$) and that of the other cesium halides (γ between 1.1 and 1.4).

TABLE VIII

TOTAL PAIR REPULSION (COLUMN 2) AND TOTAL THREE-BODY ENERGY INVOLVING NEAREST NEIGHBORS (COLUMN 3), NEXT-NEAREST NEIGHBORS (COLUMN 4), AND THIRD NEIGHBORS (COLUMN 5) FOR THE ALKALI HALIDES IN THE CESIUM CHLORIDE CONFIGURATION[a]

	Cesium chloride configuration			
Alkali halide	Pair repulsion	Three-body energy		
		$A_0C_1C_1$; $C_0A_1A_1$	$A_0A_2A_2$; $C_0C_2C_2$	$A_0A_3A_3$; $C_0C_3C_3$
LiF	24.60	−10.87	5.47	—
NaF	10.07	−7.84	0.44	—
KF	10.78	−8.53	0.82	0.30
RbF	12.60	−9.47	1.81	0.39
CsF	15.53	−10.60	3.44	0.98
LiCl	61.20	−15.12	24.91	8.05
NaCl	17.55	−11.19	4.20	1.10
KCl	11.16	−8.90	1.14	0.19
RbCl	10.75	−8.78	0.92	0.09
CsCl	10.83	−8.82	0.96	0.11
LiBr	86.41	−16.30	37.47	13.22
NaBr	22.41	−12.44	6.68	2.15
KBr	12.59	−9.58	1.95	0.49
RbBr	11.44	−9.15	1.34	0.27
CsBr	10.94	−8.91	1.06	0.11
LiI	115.02	−16.88	51.58	18.78
NaI	28.17	−13.50	9.55	3.31
KI	14.45	−10.27	2.90	0.88
RbI	12.46	−9.59	1.91	0.50
CsI	11.33	−9.11	1.29	0.24

[a] All results are expressed in units of nearest-neighbor repulsion α_1 except those in column 2 where the unit is $\alpha_1(1\text{-}b)$. The parameter b stands for the ratio between first- and second-order pair interactions for a triplet of nearest neighbors; $b \approx 0.5$ approximately.

(5) All the above conclusions are independent of α_1, the pair repulsion between nearest neighbors, and of the parameter b which measures the ratio between second-order and first-order pair interactions for a triplet of nearest neighbors.

Finally, following steps (*c*) and (*d*) of the analysis, we determine the difference in static lattice energy between the sodium chloride and the cesium chloride configurations for the alkali halides. In addition, we estimate the transition pressures for those halides which exhibit sodium chloride structure at normal pressure, following the simplified treatment given by Born and Huang (1954, Chapter 3, Eqs. 13.18 and 13.19).

In Table IX the numerical results are given for the difference in lattice energy, ΔE, between the cesium chloride and the sodium chloride structures. A *positive* value of ΔE implies that the *sodium chloride* structure is the stable one. Also included in the table are values for the transition pressure, P_t(calc.), the corresponding results of the Born–Mayer–Jacobs theory, P_t(B–M), and the experimental (Bridgman, 1952; Piermarini and Weir, 1962) values, P_t(exp.). The precise numerical values for the parameter b are not of importance for the *sign* of the difference in lattice energy. However, the transition pressures are a sensitive function of b. The values listed in Table IX give the best agreement with experimental results on transition pressures. It is important to observe that b varies regularly and in the same manner for all the alkali halide crystals.

It is seen from the results of Table IX that, by introducing three-body exchange interactions between the ions, *all the main aspects of the stability problem for alkali halide crystals can be derived on a quantitative basis.* Specifically, the theory accounts for the stability of the cesium chloride configuration for cesium chloride, bromide, and iodide. Moreover, the theory reproduces the pressure values of experimentally observed transitions from the sodium chloride to the cesium chloride configuration. Further, it is of particular interest to note that the theory agrees with recent experimental information (Piermarini and Weir, 1962) according to which rubidium fluoride is *less stable* in the sodium chloride structure than cesium fluoride.

Considering the values of the parameter b, which stands for an average ratio between second- and first-order pair interactions for a triplet of nearest neighbors, it is seen that these vary regularly and in the *same* manner for all the alkali halides; the limiting values of 0.6 and 0.4 can be explained by analogy with potential functions between rare gas atoms. To illustrate the sensitivity of the results with respect to changes in b, we find that, taking $b = 0.5$ instead of 0.6 for potassium fluoride and iodide, the values for ΔE and P_t change to 3.8 kcal/mole and 53 kbar and 2.2 kcal/mole and 13.5 kbar, respectively.

TABLE IX

DIFFERENCE IN LATTICE ENERGY, ΔE, IN KCAL/MOLE, BETWEEN THE CESIUM CHLORIDE AND THE SODIUM CHLORIDE CONFIGURATIONS FOR THE ALKALI HALIDES[a]

Alkali halide	b	ΔE (kcal/mole)	P_t(calc.) (kbar)	P_t(B–M) (kbar)	P_t(exp.) (kbar)
LiF	0.6	33.3	1,060	300	—
NaF	0.6	7.5	158	200	—
KF	0.6	6.0	83	88	—
RbF	0.4	2.2	25	68	9–15[b]
CsF	0.4	3.2	31	35	(Not observed)[b]
LiCl	0.6	14.4	221	140	—
NaCl	0.6	8.1	95	74	—
KCl	0.5	2.1	18	74	19.6
RbCl	0.5	1.05	7.7	39	4.9
CsCl	0.4	−0.6		(Not calc.)	
LiBr	0.6	11.0	140	105	—
NaBr	0.6	7.8	78	53	—
KBr	0.5	2.3	17	59	18
RbBr	0.5	0.94	6.1	30	4.5–5.0
CsBr	0.4	−0.82		(Not calc.)	
LiI	0.6	6.9	67	68	—
NaI	0.6	5.7	44	44	—
KI	0.6	2.8	17	49	17.85
RbI	0.5	0.9	4.8	22	4.0
CsI	0.4	−0.95		(Not calc.)	

[a] Also given are calculated transition pressures, P_t(calc.), in kbar, compared with the results of the Born–Mayer–Jacobs theory, P_t(B–M), and experimental values, P_t(exp.). The second column gives the values for the parameter b.

[b] See Piermarini and Weir (1962). New experiments on cesium fluoride have been announced by these authors.

The *relative magnitude* of the total three-body interactions, with respect to the crystal energy of the stable structure, lies between −1.6 and +3.3% for the fluorides, between −2.8 and +3.6% for the chlorides, between −2.2 and +3.5% for the bromides, and between −2.6 and +2.3% for the iodides. The total pair repulsion varies between 7 and 18% of the crystal energy, in good agreement with the Born–Mayer theory.

IV. Discussion of Results and Concluding Remarks

In this review we have discussed the development of the theory concerning the stability of rare gas crystals and solids of the alkali halides since 1925, when Lennard-Jones and Ingham (1925) compared the lattice energies of the face-centered cubic, body-centered cubic, and simple-cubic structures for two-atom central-type forces and when Hund (1925) investigated the stability of different ionic solids with a simple model of point-charge interactions and a repulsion between closed electron shells.

In the case of *rare gas crystals* the problem is to find the source of an extra stabilizing energy for the *face-centered cubic structure* of neon, argon, krypton, and xenon. We have seen how this extra energy can arise because of three-atom second-order interactions involving nearest neighbors of a central atom. These second-order interactions favor triangles with *large* opening θ; the essential difference between the fcc and the hcp configurations lies in the difference in three-atom interactions for three fcc triangles at $\theta = 120°$ compared to three hcp triangles at $109° 28'$.

In the case of *alkali halide crystals* the problem is to find the source of an extra stabilizing energy for the *cesium chloride configuration* of the heavy salts. Here we have seen how this extra energy arises because of three-ion first-order interactions involving nearest neighbors of a central ion. The first-order interactions favor triangles with *small* opening θ and in this respect the CsCl structure is more stable by a large margin. In rare gas solids we have to overcome a very small difference in two-atom energy; this is the reason why the cubic structure becomes more stable *as soon as* three-atom interactions start to play a role. In alkali halide crystals, on the other hand, the barrier in two-ion energy is very much higher; therefore, much larger differences in three-ion interactions are required to render the CsCl structure the stable one. For this reason the transition occurs only *at the end* of the series, with the heavy cesium halides.

A second main feature of the theory of three-body interactions, which we have discussed here, is that it concerns only *relative* effects. The reason is that relative effects are expected to be less sensitive to special characteristics of approximation procedures than absolute quantities. In addition, as we learned by compraing the Axilrod–Teller (third-order) and Rosen–Shostak (first-order) calculations, relative three-atom interactions follow the same rules, irrespective of the perturbation order. Therefore, relative interactions seem to be the most interesting and most "stable" quantities to evaluate. This circumstance excludes, of course, that we are capable of giving an explanation for stability from a minimization procedure of a theoretically evaluated *total* energy of the crystal, since two-body interactions are not known from theory with any precision. In this author's opinion it does not make much sense to impose such ultimate requirements on a theory of crystal stability.

The validity of the stability analyses which we have discussed here is tied to the *approximate* validity of an atomic description of molecular solids in terms of two-atom interactions and of a Born–Mayer model for alkali halide crystals. For both types of solids experimental evidence for this approximate validity is abundant. In the case of alkali halides, especially the results of the phenomenological treatment by Tosi and Fumi (1962) give valuable information regarding the key to the stability problem.

The next category of solids for which a similar stability analysis is expected to hold consists of the oxides, sulfides, selenides, and tellurides of beryllium, magnesium, calcium, strontium, and barium, of which the ions are again *isoelectronic* with rare gas atoms. The majority of these solids have NaCl configuration, but those with the smallest cations (Be, and Mg in MgTe) exhibit B3 and B4 structures, which are the Structure Reports symbols for, respectively, the zinc blende (sphalerite) and the wurtzite configurations. The B3 structure consists of two interpenetrating fcc lattices; in the B4 configuration these lattices are hcp. Each anion (cation) in either structure is surrounded by four cations (anions). The B3 and B4 structures are very interesting in that their difference in lattice energy for pair-potentials is extremely small. A preliminary analysis shows that three-ion interactions cause a splitting in energy between these two structures, so that we can consider also their *relative* stability. We may also expect that three-atom and three-ion interactions can give an explanation for the deviations from the Cauchy relations between their elastic constants. These calculations are at present under way.

No attempt has been made here to extend the evaluation of three-body interactions beyond second-order perturbation theory, nor to include simultaneous interactions between more than three atoms or ions. The reason is, first of all, the prohibitive complexity of higher-order calculations and, further, the belief that such a double-series expansion (in the numbers of simultaneously interacting atoms or ions on the one hand, and in the order of perturbation theory on the other hand) must converge rapidly in order to render an "atomic description" of these solids at all valid. The comparison with the Axilrod-Teller third-order effect (Fig. 3) supports this expectation.

REFERENCES

ALLEN, L. C. (1962). *Science* **138**, 892.
ALLEN, L. C. (1963). *Nature* **197**, 897.
AXILROD, B. M. (1949). *J. Chem. Phys.* **17**, 1349.
AXILROD, B. M. (1951a). *J. Chem. Phys.* **19**, 719.
AXILROD, B. M. (1951b). *J. Chem. Phys.* **19**, 724.
AXILROD, B. M., and TELLER, E. (1943). *J. Chem. Phys.* **11**, 299.
AYRES, R. V., and TREDGOLD, R. H. (1955). *Phys. Rev.* **100**, 1257.
AYRES, R. V., and TREDGOLD, R. H. (1956). *Proc. Phys. Soc.* (*London*) **B69**, 840.
BADE, W. L. (1957). *J. Chem. Phys.* **27**, 1280.
BADE, W. L. (1958). *J. Chem. Phys.* **28**, 282.
BADE, W. L., and KIRKWOOD, J. G. (1957). *J. Chem. Phys.* **27**, 1083.
BALLHAUSEN, C. J. (1962). "Introduction to Ligand Field Theory." McGraw-Hill, New York.
BARRETT, C. S., and MEYER, L. (1964). *J. Chem. Phys.* **41**, 1078.
BARRON, T. H. K., and DOMB, C. (1955). *Proc. Roy. Soc.* **A227**, 447.
BEGBIE, G. H., and BORN, M. (1947). *Proc. Roy. Soc.* **A188**, 179.
BLEICK, W. E., and MAYER, J. E. (1934). *J. Chem. Phys.* **2**, 252.
BORN, M. (1940). *Proc. Cambridge Phil. Soc.* **36**, 160.
BORN, M. (1942). *Proc. Cambridge Phil. Soc.* **38**, 82.
BORN, M. (1944). *Proc. Cambridge Phil. Soc.* **40**, 262.
BORN, M., and HUANG, K. (1954). "Dynamical Theory of Crystal Lattices." Oxford Univ. Press, London and New York.
BORN, M., and MAYER, J. E. (1932). *Z. Physik* **75**, 1.
BOYS, S. F. (1950). *Proc. Roy. Soc.* **A200**, 542.
BRIDGMAN, P. W. (1952). "The Physics of High Pressure." G. Bell, London.
BROWN, F. W. (1933). *Phys. Rev.* **44**, 214.
COCHRAN, W. (1959). *Proc. Roy. Soc.* **A253**, 260.
COLWELL, J. F. (1960). Thesis, Cornell University (unpublished).
CORNER, J. (1939). *Trans. Faraday Soc.* **35**, 711.
CUTHBERT, J., and LINNETT, J. W. (1958). *Trans. Faraday Soc.* **54**, 617.
DANON, F. (1965). To be published.
DE WETTE, F. W. (1959). *Physica* **25**, 1225.
DICK, B. G. (1963). *Phys. Rev.* **129**, 1583.
DICK, B. G., and OVERHAUSER, A. W. (1958). *Phys. Rev.* **112**, 90.
DOBBS, E. R., and JONES, G. D. (1957). *Rept. Progr. Phys.* **20**, 516.
DOMB, C., and SALTER, L. (1952). *Phil. Mag.* [7] **43**, 1083.
DONIACH, S. (1963). *Phil. Mag.* [8] **8**, 129.
DUGDALE, J. S., and SIMON, F. E. (1953). *Proc. Roy. Soc.* **A218**, 291.
EVDOKIMOVA, V. V., and VERESHCHAGIN, L. F. (1963a). *Soviet Phys.—Solid State* **4**, 1438.
EVDOKIMOVA, V. V., and VERESHCHAGIN, L. F. (1963b). *Soviet Phys.—JETP* **16**, 855.
FOREMAN, A. J. E., and LIDIARD, A. B. (1963). *Phil. Mag.* [8] **8**, 97.
GOEPPERT-MAYER, M., and KANE, G. (1940). *J. Chem. Phys.* **8**, 642.

GRABER, H. W., and PRESENT, R. D. (1962). *Phys. Rev. Letters* **9**, 247.

HANLON, J. E., and LAWSON, A. W. (1959). *Phys. Rev.* **113**, 472.

HIRSCHFELDER, J. O., and RICE, W. E. (1954). *J. Chem. Phys.* **22**, 187.

HIRSCHFELDER, J. O., CURTISS, C. F., and BIRD, R. B. (1954). "Molecular Theory of Gases and Liquids." Wiley, New York.

HUGGINS, M. L., and MAYER, J. E. (1933). *J. Chem. Phys.* **1**, 643.

HUND, F. (1925). *Z. Physik* **34**, 833.

JACOBS, R. B. (1938). *Phys. Rev.* **54**, 468.

JANSEN, L. (1957). *Physica* **23**, 599.

JANSEN, L. (1958). *Phys. Rev.* **110**, 661.

JANSEN, L. (1962). *Phys. Rev.* **125**, 1798.

JANSEN, L. (1963a). *Phys. Letters* **4**, 91.

JANSEN, L. (1963b). *Phil. Mag.* [8] **8**, 1305.

JANSEN, L. (1964). *Phys. Rev.* **135**, A1292.

JANSEN, L., and DAWSON, J. M. (1954). *J. Chem. Phys.* **22**, 1619.

JANSEN, L., and DAWSON, J. M. (1955). *J. Chem. Phys.* **23**, 482.

JANSEN, L., and McGINNIES, R. T. (1956). *Phys. Rev.* **104**, 961.

JANSEN, L., and ZIMERING, S. (1963). *Phys. Letters* **4**, 95.

JORTNER, J., RICE, S. A., and WILSON, E. G. (1963). *J. Chem. Phys.* **38**, 2302.

KEESOM, W. H., and TACONIS, K. W. (1938). *Physica* **5**, 161.

KESTNER, N. R., and SINANOĞLU, O. (1963). *J. Chem. Phys.* **38**, 1730.

KIHARA, T. (1953). *Rev. Mod. Phys.* **25**, 831.

KIHARA, T. (1958). *Advan. Chem. Phys.* **1**, 267.

KIHARA, T. (1960). *J. Phys. Soc. Japan* **15**, 1920.

KIHARA, T. (1963a). *Advan. Chem. Phys.* **5**, 147.

KIHARA, T. (1963b). *Acta Cryst.* **16**, 1119.

KIHARA, T., and KOBA, S. (1952). *J. Phys. Soc. Japan* **7**, 348.

KIMEL, S., RON, A., and HORNIG, D. F. (1964). *J. Chem. Phys.* **40**, 3351.

KITTEL, C. (1957). "Introduction to Solid State Physics." Wiley, New York.

KNOX, R. S., and REILLY, M. H. (1964). *Phys. Rev.* **135**, A166.

LENNARD-JONES, J. E., and INGHAM, A. E. (1925). *Proc. Roy. Soc.* **A107**, 636.

LINNETT, J. W., and POË, A. J. (1951). *Trans. Faraday Soc.* **47**, 1033.

LOMBARDI, E., and JANSEN, L. (1964a). *Phys. Rev.* **136**, A1011.

LOMBARDI, E., and JANSEN, L. (1964b). *Phys. Rev. Letters* **12**, 11.

LÖWDIN, P.-O. (1948). "A Theoretical Investigation into Some Properties of Ionic Crystals." Almquist & Wiksell, Uppsala.

LÖWDIN, P.-O. (1956). *Advan. Phys.* (*Phil. Mag. Suppl.*) **5**, 1.

LONDON, F. (1930a). *Z. Physik Chem.* **B11**, 222.

LONDON, F. (1930b). *Z. Physik* **63**, 245.

McGINNIES, R. T., and JANSEN, L. (1956). *Phys. Rev.* **101**, 1301.

McLACHLAN, A. D. (1963). *Mol. Phys.* **6**, 423.

MARGENAU, H. (1939). *Rev. Mod. Phys.* **11**, 1.

MAY, A. (1937). *Phys. Rev.* **52**, 339.

MAY, A. (1938). *Phys. Rev.* **54**, 629.

MAYER, J. E. (1933). *J. Chem. Phys.* **1**, 270.

MEYER, L., BARRETT, C. S., and HAASEN, P. (1964). *J. Chem. Phys.* **40**, 2744.

MIDZUNO, Y., and KIHARA, T. (1956). *J. Phys. Soc. Japan* **11**, 1045.

MISRA, R. D. (1940). *Proc. Cambridge Phil. Soc.* **36**, 173.

MULLIKEN, R. S. (1951). *In* "Conference on Quantum Mechanical Methods in Valence Theory," p.108. Office of Naval Research, Washington, D.C.

MUNN, R. J. (1964). *J. Chem. Phys.* **40**, 1439.

MUNN, R. J. (1964). Private communication.

MUTO, Y. (1943). *Proc. Phys. Math. Soc. Japan* **17**, 629.

NABARRO, F. R. N., and VARLEY, J. H. O. (1952). *Proc. Cambridge Phil. Soc.* **48**, 316.

NARDELLI, G. R., and REPANAI CHIAROTTI, A. (1960). *Nuovo Cimento* [10] **18**, 1053.

PAULING, L. (1960). "The Nature of the Chemical Bond," 3rd ed. Cornell Univ. Press, Ithaca, New York.

PIERMARINI, G. J., and WEIR, C. E. (1962). *J. Chem. Phys.* **37**, 1887.

PITZER, K. S. (1963). *Science* **139**, 414.

POLLACK, G. L. (1964). *Rev. Mod. Phys.* **36**, 748.

POWER, S. C. (1942). *Proc. Cambridge Phil. Soc.* **38**, 62.

PRINS, J. A. DUMORÉ, J. M., and TJOAN, L. T. (1952). *Physica* **18**, 307.

ROSEN, P. (1953). *J. Chem. Phys.* **21**, 1007.

RUNDLE, R. E. (1963). *J. Am. Chem. Soc.* **85**, 112.

SCHALLAMACH, A. (1939). *Proc. Roy. Soc.* **A171**, 569.

SCHUCH, A. F., OVERTON, W. C., JR., and BROUT, R. (1963). *Phys. Rev. Letters* **10**, 429.

SHAVITT, I. (1963). *In* "Methods in Computational Physics" (B. Alder and S. Fernbach, eds.), Vol. 2, p.1. Academic Press, New York.

SHERWOOD, A. E., and PRAUSNITZ, J. M. (1964). *J. Chem. Phys.* **41**, 413 and 429.

SHOSTAK, A. (1955). *J. Chem. Phys.* **23**, 1808.

SINANOĞLU, O., and PITZER, K. S. (1960). *J. Chem. Phys.* **32**, 1279.

SPARNAAY, M. J. (1959). *Physica* **25**, 217.

TOSI, M. P., and FUMI, F. G. (1962). *J. Phys. Chem. Solids* **23**, 359.

VARSHNI, Y. P., and SHUKLA, R. C. (1961). *J. Chem. Phys.* **35**, 582.

WALLACE, D. C. (1964). *Phys. Rev.* **133**, A153.

WOJTALA, J. (1964). *Acta Phys. Polon.* **25**, 27.

WOODS, A. D. B., COCHRAN, W., and BROCKHOUSE, B. N. (1960). *Phys. Rev.* **119**, 980

ZIMERING, S. (1965). *J. Math. Phys.* **6**, 336.

Charge Fluctuation Interactions in Molecular Biophysics

HERBERT JEHLE

Physics Department
The George Washington University
Washington, D.C.

I. Introduction

Molecular biophysics is confronted with the question as to the types of intermolecular interactions on various levels of biological organization. Particular interest into biologically specific types of interactions naturally leads us to the question of interaction of biological macromolecules with each other. These exhibit their biological specificity often through the interactions of their side chains (amino acid side chains, nucleotide side chains, or other groups) with the side chains of other macromolecules. It is to such interactions, though not exclusively, that we want to turn our special attention.

A clarification of the role of intermolecular forces in biology is a basic issue for any understanding of biophysical processes. It is also a puzzling issue because the interplay between the different kinds of forces is such an intricate one. Often, it is only barely possible to delineate and experimentally establish where one kind of force, and where another, plays the predominant role in a given process.

This situation may be clarified if we gain a more complete understanding of the theoretical aspects of the various types of forces. One would naturally like to establish from theory some clues that would be of help in deciding, by experiment, what is the involvement of the different forces: there are many measurable optical effects and thermodynamic properties which follow from the theory of intermolecular forces.

Apart from this approach, one sometimes finds that the nature of the specific biological phenomenon narrows down the possible types of interactions. Strangely enough, it is often the complexity of a given phenomenon that helps to establish the choice of the appropriate mechanism responsible for it. This is seen in particular in interactions involving complementarity of molecular structure and in those which are characterized by self-recognition of similar molecules.

It is not our intention to go into a detailed review of well-known intermolecular forces. The covalent bond (Pauling, 1960; Pauling and Wilson, 1935) depends on the sharing of electron pairs and comes about only if the electronic shells of the interacting molecules overlap (Van Vleck and Sherman, 1934; Heitler and F. London, see Heitler, 1945; Pullman and Pullman, 1963; Slater 1963; Eyring *et al.*, 1944). The ionic bond depends on opposite polarity of certain groups of the interacting molecules. Also the hydrogen bond is closely related to this (Pimentel and McClellan, 1960; Doty *et al.*, 1960). The charge transfer force depends on a short-time exchange of charge between the interacting molecules (Mulliken and Person, 1962). The van der Waals forces depend on polarization and on polarizabilities of the molecules; the last discovered van der Waals force (by London, R. Eisenschitz, and S. C. Wang, see London, 1936, 1942) is concerned with the important attraction between two merely polarizable molecules (Hirschfelder, 1965). These are called charge fluctuation forces.

To complicate matters further, one has to take into consideration the involvement of the medium in which the interacting molecules are immersed (Sinanoğlu *et al.*, 1964; Kestner and Sinanoğlu, 1963).

II. Free Energy of Interaction

We shall now discuss one of the many types of intermolecular interactions which are encountered in biophysics, the charge fluctuation interactions. Reference is made to other types of interactions only to the extent of clarification.

The molecules with whose interaction the biophysicist usually has to deal are not in a setting of two isolated molecules (such as individual molecules in collisions occurring in a rarified gas). They are rather to be considered as imbedded in a surrounding, often aqueous, which hereafter is referred to as the "medium."

We speak of intermolecular "interactions" rather than of intermolecular "forces" because the more general term "interaction" not only includes the forces (energies) involved, but refers to the entire setting which includes the order of arrangement of the molecules, their mutual orientation, and the kind of a surrounding medium in which the molecules are placed.

The word interaction is frequently used as a synonym for "interaction energy." The interaction "force" is the gradient of the interaction energy; that is, the negative of the interaction energy is the work contributed by an attractive interaction force if the pair of molecules, starting off with infinite separation R, are permitted to approach each other.

If one considers a pair of free molecules which has no heat to exchange with a nonexisting medium, it is the internal energy of the pair of interacting molecules which determines that attractive force. If, on the other hand, this pair of molecules is imbedded in a temperature bath (a medium of given temperature), it is the Helmholtz free energy A of the pair of molecules which determines the attractive force. A heat bath of constant temperature ($\Delta T = 0$) is able to supply heat energy to the pair of molecules imbedded in it, maintaining statistically a Boltzmann distribution over the energy levels of the molecule pair. This implies that the work contributed by the attractive force between the two molecules ($-\Delta A > 0$) is not limited to the loss of internal energy ($-\Delta E > 0$) which they originally had, but that replenishment of energy from the heat bath (heat input $T\,\Delta S > 0$) gives an entropy contribution. In other words, it is the change of Helmholtz free energy with intermolecular distance

$$\frac{\Delta A}{\Delta R} = \frac{\Delta(E - TS)}{\Delta R} = \frac{\Delta E - T\,\Delta S}{\Delta R} \tag{1}$$

which determines the attractive force. In the language of statistical mechanics this means

$$\sum_i n_i \frac{\Delta \epsilon_i}{\Delta R} = \sum_i \frac{\Delta(n_i \epsilon_i)}{\Delta R} - \sum_i \epsilon_i \frac{\Delta n_i}{\Delta R} \tag{2}$$

where n_i is the statistical occupancy of the ith energy state ϵ_i.

When formulating the interaction between molecules it is necessary to do so on the basis of a mathematical model scheme. Actually, as one describes any type of interaction, one defines it through a usually simplified model depicting that type of interaction.

Such model schemes have often been simply postulated; a proper justification can be given to such schemes only if the theory is complete, encompassing the several types of interactions which come to play a role in a particular instance. Simplified model schemes are still very useful as a means to provide a discussable picture of a complex situation, and to permit confrontation of a theory with a broad range of diversified data. In this manner covalent bonding, ionic bonding, hydrogen bonding, charge transfer interactions, and van der Waals interactions can be discussed. The possibility of defining such different types of interactions is based on the assumption of mutually independent model schemes which, when considered together, are supposed to conform to the general theory. When considered separately, these models are simple enough to make qualitative discussions possible.

III. Interactions between Two Molecules Held at Given Temperature

The two molecules under discussion may be thought of as imbedded in a medium. The medium may at this point be considered only as providing a bath of given temperature. All other involvement of the medium in the interaction is excluded from the model with which this chapter deals.

When the molecules are near each other, hydrogen bridges and other bridges might form. Possibly an overlap of the electronic shells of the two molecules might occur, leading to well-known consequences such as close range repulsion between molecules, covalent bond formation (Heitler–London theory, see Heitler, 1945), and electronic charge transfer between molecules (Mulliken and Person, 1962; Briegleb, 1961). Even though such charge transfer may be of a fluctuating type, e.g., when the interacting molecules are imbedded in a medium, we may use the special terminology "charge transfer" for this mechanism.

At large intermolecular distance R, overlap of electronic shells becomes negligible. It is then only through the intermediary of the electromagnetic field between the molecules that interaction occurs. Such interactions may be described as a photon exchange, virtual exchange of two photons and similar effects.

Let us first consider the particular interaction of one molecule in a definite excited electronic energy state and another in an unexcited state (rather than a statistical average of interactions at random thermal excitations.) One may then have a "first-order" interaction through the transmission of a photon. The probability of this occurring is very strongly dependent on resonance existing between one molecule and the other. This occurs if some transition frequencies of the two interacting molecules are identical.

Considering second the interaction between a pair of molecules involving an interchange of two photons, "second-order" effects arise which are important but which do not have such a sharp resonance response as the aforementioned first-order resonance effects, or as a radio receiver tuned to the frequency of a single station.

It is these "second-order" effects which we wish to discuss in more detail. These are frequently called "van der Waals forces" and may themselves be subdivided into several types (cf. London, 1936).

Historically, the first interpretation of van der Waals attraction was given by Keesom in 1912. His "orientation effect" suggested that interaction of freely rotating electrically polar molecules, having a dipole moment of fixed magnitude, with another such molecule may occur due to the statistical preference of the dipoles for orientations that are energetically advantageous. As the preferential electric molecular orientations are countereffected by the randomization tendency of Brownian motion, this type of interaction is temperature-dependent. As there exist important van der Waals interactions which are temperature-independent and in which both molecules do not necessarily possess fixed dipole moments, Keesom's theory had to be supplemented by other explanations.

In 1920, Debye and H. Falkenhagen suggested that a molecule having a fixed dipole moment might interact with a molecule capable of acquiring a dipole moment when placed in an electric field of dipolar molecules (see Debye, 1920). This was called the "induction effect." A molecule is called polarizable if it acquires a dipole moment when placed in an electric field. That interaction is temperature-independent, but still depends on molecules which are inherently polar, and thus cannot account for all types of van der Waals interactions. Besides, contrary to the properties of this mechanism of intermolecular interaction, experiments show that van der Waals interactions of two molecules, I and II, with a third one, III, are in many instances independent of whether or not I and II are interacting with each other. (In these instances there is also no evidence of saturation of pair-interactions as is the case in covalent bonding.) Finally, both the orientation and the induction effects lead to interaction energies frequently much below the observed values.

In 1930, London, Eisenschitz, and Wang suggested a mechanism involving an interaction between the fluctuating electronic charge distribution in one molecule and the fluctuating charge distribution in another molecule (see London, 1936). This "dispersion effect" is an interaction between two polarizable molecules without permanent dipole moments.

The term "charge fluctuations" was coined by London and has been maintained by theoretical physicists to characterize the situation. The equivalent term "dispersion forces" has also been used for these forces resulting from charge fluctuation interactions between two atoms or two molecules. This term arises simply because the theory of optical dispersion of media implies similar mechanisms. This second-order perturbation theory involves atomic or molecular electrons in their response to an oscillating electric field (Hirschfelder *et al.*, 1954; Hirschfelder 1941, 1964, 1965; Hirschfelder and Eliason, 1957; Löwdin, 1955, 1956, 1957, 1960, 1962, 1963; Linderberg and Shull, 1960; Linderberg, 1962; Feynman, 1939; Hellmann, 1937; Margenau, 1939; Förster, 1951, 1963).

Still another term frequently used is "correlations." This term, however, refers usually to the case of intra-atomic interactions of charge fluctuations caused by the various electrons of that atom. "Correlation energy" is used still more specifically as the difference between the actual experimentally observed energy and the one calculated by what is called the Hartree–Fock method. This difference is essentially brought about by correlations of electronic motions. Such correlation occurs whenever there are charge fluctuations, in the atomic case fluctuations between the motions of different electrons in an atom (Wigner, 1934, 1938; Pines, 1955; Löwdin, 1962; Tomonaga, 1950, 1955).

The model we have considered implies absence of overlap of electronic shells. It thus disregards charge transfer interactions and other important effects. A complete theory implying in a single calculation van der Waals interactions at large distances and the short-distance interactions due to overlap of electronic shells has recently been made by Sinanoğlu, *et al.* (1964).

Regarding the large-distance (R) London interactions, it is important to note that, with no electronic overlap or other possibility of charge transfer, the total charge on each of the two molecules remains fixed and does not fluctuate. Fluctuations occur in the distribution of charge inside each of the two molecules. Dipolar terms of that fluctuation, i.e., dipolar terms of polarizability, are not only the most important ones usually, but also the ones that permit fairly simple calculations and subsequent discussion. We will therefore confine our discussion to this type of interaction mechanism.

There are two "limit cases," the quantum and the classical limit, of the London charge fluctuation interactions between two molecules. Both cases can be most conveniently discussed in terms of the Drude oscillator model which registers the response of the molecules to an oscillatory electric field in terms of an oscillatory displacement of electric charge distribution of the responding molecules.

The "quantum limit" refers to the situation when the oscillator frequencies ω_l are so high that excited states [of energies $(n_l + \frac{1}{2})\hbar\omega_l$] of these oscillator motions are out of reach by thermal motion; only the nonexcited ground state is occupied. The quantum mechanical ground state (of energy $\frac{1}{2}\hbar\omega_l$), unlike the classical one, implies fluctuation of electronic motion, as does every other quantum state. This is simple to understand. If one makes a classical analog of an atomic or molecular stationary state in quantum mechanics, commonly called a "quantum state," one has to represent it by a statistical distribution of electron orbits. Physical quantities like electric dipole moment (polarization) of the atom or molecule under consideration have a certain expectation value (often zero) in that quantum state, and a certain fluctuation as every statistical theory implies.

The fluctuations of charge distribution in one molecule, I, and those in the other molecule, II, interact so as to give preference to the energetically more advantageous fluctuations; this causes an attraction between the molecules. In the Drude model this preferential correlation of fluctuations is taken into account by calculation of the effect of interaction, for the inphase mode (synchronous inphase oscillation of the two oscillators) and for the antiphase mode. Both are participating in the ground state. Even though, in first order, the effects of electrical interaction of inphase mode and that of antiphase mode cancel out (London, 1936), the second-order terms give a net effect of attraction. The energy or free energy of attraction turns out to be

$$-\Delta A = \tfrac{3}{2}R^{-6}\alpha_{\mathrm{I}}\alpha_{\mathrm{II}}\hbar\,\frac{\tilde{\omega}_{\mathrm{I}}\tilde{\omega}_{\mathrm{II}}}{\tilde{\omega}_{\mathrm{I}} + \tilde{\omega}_{\mathrm{II}}} \tag{3}$$

for two molecules, each represented by a single isotropic harmonic oscillator, of the polarizabilities α_l and frequencies $\tilde{\omega}_l$, and $\Delta A = A_R - A_\infty$. $\tilde{\omega}_l$ denotes the oscillator frequencies when the distance between the two molecules is $R = \infty$ whereas ω_l represents the normal mode frequencies of coupled molecule oscillations where R is finite.

In the "classical limit" one considers the case in which the classical oscillator frequencies $\tilde{\omega}_l$ are so low that energy steps $\hbar\tilde{\omega}_l$ between successive quantum levels are small compared with the average energy kT of thermal motion. In this case one cannot expect the electrons to remain in a definite quantum state. Accordingly, the electrons perform a Brownian motion over the thermally accessible quantum states. Each quantum state of the interacting pair of molecules makes, through its charge fluctuations, a contribution to the attraction. These various contributions are added with a statistical weighting factor

$$n_j/\sum_i n_i = \exp(-\epsilon_j/kT)/\sum_i \exp(-\epsilon_i/kT) \tag{4}$$

which indicates the fractional occupancy $n_j/\sum_i n_i$ for the state j of the molecule pair; ϵ_j is the energy of the quantum state. With this, the attractive force between a molecule pair can be shown to become

$$\begin{aligned}\mathscr{F} &= \sum_j n_j \frac{\partial \epsilon_j}{\partial R}\bigg/\sum_i n_i = \left(\frac{\partial A}{\partial R}\right)_T \\ &= \frac{\partial}{\partial R}\left[-kT \ln \sum_j \exp(-\epsilon_j/kT)\right]. \end{aligned} \tag{5}$$

A simple oscillator calculation shows that this implies

$$-\Delta A = 3R^{-6}\alpha_{\mathrm{I}}\alpha_{\mathrm{II}}kT \tag{6}$$

in the classical limit, in other words, a formula similar to the former quantum mechanical London formula. In (6), however, kT, the energy of Brownian motion, rather than the zero-point energy $\frac{1}{2}\hbar\tilde{\omega}$ in (3), characterizes the energy of fluctuation which underlies the interaction free energy $-\Delta A$. Between these quantum and classical limits, $\frac{3}{4}\hbar\tilde{\omega}$ and $3kT$, respectively, the formula

$$-\Delta A = \tfrac{3}{2}kT\left\{\frac{\hbar\tilde{\omega}}{2kT}\coth\frac{\hbar\tilde{\omega}}{2kT} + \left(\frac{\hbar\tilde{\omega}}{2kT}\right)^2\bigg/\sinh^2\frac{\hbar\tilde{\omega}}{2kT}\right\}\frac{\alpha^2}{R^6} \tag{7}$$

gives the single oscillator energy if the two isotropic oscillators have the same frequency $\tilde{\omega} = \tilde{\omega}_{\mathrm{I}} = \tilde{\omega}_{\mathrm{II}}$.

In the actual interactions of molecules a more general situation is involved: many oscillators whose frequencies and orientations are widely diversified are needed to adequately represent a molecule in its interaction properties. Several types of solutions have been found to this problem (London, 1936; Yos, Bade, Jehle, see Yos, 1958; McLachlan, 1963; Salem, 1964; Linder, 1964). It can be shown that the interaction energy is of the form

$$-\Delta A = (\hbar/4\pi i R^6)\int_{-\infty}^{+\infty} a_{\mathrm{I}}(\omega)a_{\mathrm{II}}(\omega)\,d\omega$$

where $a_{\mathrm{I}}\,a_{\mathrm{II}}$ represents the trace of the product of the dynamic polarizability interaction tensors of molecules I and II taken on the imaginary axis of the complex ω plane. Spelled out in more detail

$$-\Delta A = \tfrac{1}{2}kT\sum_{S=-\infty}^{+\infty}\;\sum\sum_{\mu,\nu=1}^{3} W_{S\mu\nu_{\mathrm{I}}}W_{S\mu\nu_{\mathrm{II}}}$$

where, in the Drude model,

$$W_{S\mu\nu_{\mathrm{I}}} = -R^{-3}\alpha_{\mu\nu_{\mathrm{I}}} = -R^{-3}\sum_{l=1}^{N_{\mathrm{I}}}[1 + (S2\pi kT/\hbar\tilde{\omega}_l)^2]^{-1}\sum_{\rho,\sigma=1}^{3} v_{\mu\rho}\alpha_{l\rho\sigma}v_{\sigma\nu}$$

and where

$$v_{\rho\sigma} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix}, \qquad \omega = iS2\pi kT/\hbar;$$

the polarizibility tensor $\alpha_{l\rho\sigma}$ of the oscillator l of frequency $\tilde{\omega}_l$ contributes to the polarizability interaction tensor $\alpha_{\mu\upsilon\mathrm{I}} = \alpha_{\mathrm{I}}$ through multiplication with the dipole standard matrices $v_{\rho\sigma}$ and through the above summation over the Drude linear oscillators l of the molecule I, and similarly for molecule II.

London has shown that all types of van der Waals forces, those due to the orientation effect, those due to the induction effect, and those due to the dispersion effect, can all be systematically covered by a single quantum mechanical calculation. The orientation and the induction effects are accounted for by including in the calculations a proper consideration of the quantum states of rotational motion of the molecules.

When molecules of the size of a nucleotide base, or of a large amino acid side chain, come to *lie next to a similar molecule*, the interaction energy might be of the order of several kilocalories per mole. It is important to note that the London interaction energy is proportional to $\omega\alpha^2$, i.e., proportional to f^2/ω^3 (where f is the oscillator strength of the representative oscillators); this means that the region of the spectrum which contributes most to the interaction energy is in the extreme ultraviolet, way beyond 2000 Å. To get an adequate estimate of the size of the London–van der Waals forces, one can use the data of London (1936), of Rabinowitch and Epstein (1941), of Tinoco and DeVoe (1962), and of Platzman (1962). Platzman's pioneering work, thus far done for molecules while in the gas phase, permits the evaluation of f values, and thus of polarizabilities for the entire important frequency range from the ultraviolet to the soft X-ray region, by a judicious combination of accurate absorption and dispersion data in the accessible region of the spectrum with the use of the Kramers–Kronig (1928) dispersion formulae and with the basic Thomas–Reiche–Kuhn sum rule.

It should be remarked that, when one looks into realms beyond biophysics, there are important types of charge fluctuation interactions. The Russian school has calculated and experimented with van der Waals forces between macroscopic bodies. They use ingenious methods from the general theory of fluctuations, based on the Callen–Welton theorem (Callen and Welton, 1951; Lifshitz and co-workers, see Lifshitz, 1955; Landau and Lifshitz, 1958, 1960; Abrikosov *et al.*, 1959; Derjaguin *et al.*, 1956; Becker, 1955; Jehle *et al.*, 1964b), to calculate the attractive forces between matter in bulk, and they brought theory and experiment to match. There is also the wide and promising field of van der Waals crystals which will be greatly influenced by progress in the understanding of charge fluctuation interactions.

IV. Participation of the Surrounding Aqueous Medium in the Interaction between Molecules

(Scatchard, 1931, 1937, 1952; Debye and Hückel, 1923; Onsager, 1936)

As before, our intent is to discuss only the charge fluctuation aspect of intermolecular interactions in biophysics. When the pair of molecules under discussion is imbedded in an aqueous surrounding, the protons of the medium, while not entirely free to leave the immediate surrounding of one of the molecules, may move about the surface of that molecule or make fluctuating movements of any kind. In this way, they might migrate from one to the other molecule or into the medium and thereby cause a protonic charge transfer, an actual fluctuation of net charge on a molecule. As before, in the case of electronic charge transfer, this case might be dealt with separately. In migrating over a molecule's surface (due to Brownian motion, temperature motion), the protons give rise to fluctuating dipole, quadrupole, etc., moments. These interact with similar fluctuations of charge distribution over the other molecule and thus lead to an attraction by a mechanism which is in many respects analogous to London's electronic charge fluctuations. Kirkwood and Shumaker (1952) discovered that mechanism and it has been recently re-examined by Oncley and Scheider (1964).

Perhaps the biggest effect among protonic types of charge fluctuations might be due to structuralized regions of water, "hydrotactoids," surrounding the molecules under consideration, in particular polyelectrolytes. Such hydrotactoids consist of groups of tens or hundreds of water molecules with their covalent OH bonds oriented, more or less, in one direction, their hydrogen bonds in the other direction. Large-magnitude charge fluctuations might occur in any hydrotactoid due to an almost synchronous flipping back and forth of the protons over their hydrogen bridges.

It might be remarked that these protonic and electronic charge fluctuations give rise to an attraction that is often counterbalanced by either the short-range "van der Waals repulsion" (de Boer and Heller, 1937) or the electronic repulsion due to similar charges on the molecules which are only partly compensated by Bjerrum defects, by attachment of small gegenions, and by a Debye–Hückel–Onsager layer (Debye and Hückel, 1923; Onsager, 1933, 1936) of neutralized gegenions (Vinograd, 1935). Hamaker (1938, 1948) has proposed that the stability of lyophobic colloids might be understood in this manner (Overbeek, 1948, 1952).

A similar mechanism may also lead to an understanding of the equilibrium distance between regularly spaced fibers found in several kinds of biological ultrastructures. Indeed, if the fibers were to come too closely together, the hydrotactoids would be squeezed out, resulting in a disappearance of the attraction mechanism which keeps the fibers together in bundles. This would, in turn, lead to a pushing apart of the fibers by the short-range and electrostatic repulsions.

There are, however, several other important interactions coming into the picture. When two side chains on larger molecules approach each other, there will be considerable changes in the possible associations of bivalent ions forming bridges between them, and, of course, changes in hydrogen bonding with the surrounding aqueous medium (Doty *et al.*, 1960), depending on the molecules being polyelectrolytes or apolar. Also there will be a most important entropy change, in particular in situations where apolar bonding (Nemethy and Scheraga, 1963) occurs.

V. Charge Fluctuations and Apolar Bonding

Apolar bonding is, in terms of intermolecular interactions, a composite phenomenon. It indicates, e.g., in the case of amino acid side chains of a globular protein, the tendency of nonpolar side chains to huddle together, pointing toward the inside of the globule rather than into the surrounding aqueous medium. An apolar bond comes about when two or more nonpolar groups form such an arrangement, thereby decreasing the extent of contact with the surrounding water. This effect is to be understood in terms of several types of interaction.

First, the immersion of a nonpolar side chain into aqueous surroundings is often accompanied by a small negative enthalpy change (compared with its immersion into an organic solvent). It may be said with Klotz (1962) that in this respect at least apolar molecules do not fear water.

Second, a heightened contact of an apolar molecule, or, in particular, of an apolar side chain with the surrounding water is accompanied by a strongly negative change in entropy, dominating so much over the first effect that there results a large positive free energy of solution, or of contact of the apolar side chain with water.

These two effects may be understood in terms of an increase of ordering, or of structuralization of water near the apolar side chain if that side chain reaches towards water (Frank and Wen, 1957; Frank, 1958; Kauzmann, 1954; Nemethy and Scheraga, 1963; Tanford, 1961). Accordingly, the nonpolar groups have a tendency to huddle into the inside of a globular molecule.

Third, while several such side chains huddle into the inside and thus achieve there some entropy gain, they might reach further stabilization because of van der Waals interactions thereby achieved. In this manner, charge fluctuation interactions are also to some degree involved in apolar bonding.

VI. Charge Fluctuation Interactions and Biological Specificity

Biologically specific interactions are usually discussed in terms of *complementarity* of distribution of charge over the corresponding contact points of the interacting molecules. A particularly well-known example of this is the complementarity between the two halves of the double-stranded nucleic acid helix. Naturally, such specific interactions, in order to become possible, demand also some structural, i.e., steric, complementarity between the interacting partners.

Steric complementarity of the interacting molecules may, however, by itself, provide for specific interaction. Pauling (1940, 1948, 1957; Pauling and Delbrück, 1940; Pauling *et al.*, 1943) calculated the contribution of van der Waals forces to this effect; the mere absence of empty interspaces between interacting molecules certainly leads to stability as there is a general van der Waals attraction between all atoms.

Another type of specific interaction arises when London–van der Waals interactions, as first pointed out by Hamaker (1937), aided by charge fluctuation interactions in general cause preferential association of identical molecules as nearest neighbors. That effect may account for *self-recognition.* This specific charge fluctuation interaction may come into prominence when static charge distributions of the interacting molecules are either negligible or compensated by gegenions. We consider the close-range interaction of a pair, for example, of amino acid side chains or of nucleotides when they are in nearest neighbor positions, particularly when aromatic side chains pair up side by side. And we consider specific interactions of two protein molecules when several pairs of corresponding side chains of this protein pair interact. If the two proteins are identical, many side chains of one protein molecule may interact with the exactly corresponding set of side chains of the other molecule. The side chains may interleave somewhat in the fashion of a zipper where the side chains are sticking out along two corresponding edges of two identical structured proteins (e.g., along two corresponding α helixes); the structured proteins might even otherwise be different. Or, the side chains may belong to one or to several branches of an open-chain-like molecule and these side chains may pair up with those of a second such long molecule which is identical with the first one. For mucopolysaccharides and other chain molecules a similar effect may be apt to occur. Pictures of models of both types of interactions are given in the BioPolymer Symposium #1 on Polypeptides and Polynucleotides by Jehle *et al.* (1964c) and also in the Proceedings of the National Academy of Sciences (see Jehle *et al.*, 1957, 1963).

The theory of this specific interaction which provides for a self-recognition of a molecule (e.g., an amino acid or a nucleotide side chain) by *an identical partner molecule is essentially simple.* Consider a sequence of molecules $\textcircled{I}\textcircled{I}\textcircled{I}\textcircled{II}\textcircled{I}\textcircled{II}\textcircled{I}\textcircled{I}$ where $\textcircled{I}$ and $\textcircled{II}$ represent two different single molecules of similar size, or $\textcircled{I}$ might actually stand for a group of medium molecules lumped together into an aggregate of same size as the molecule type $\textcircled{II}$. *Denote the interaction energy of neighboring pairs by* $\Delta A_{\mathrm{I,II}}$, $\Delta A_{\mathrm{I,I}}$, respectively, and neglect second nearest neighbor interactions. Then the interaction energy of that arrangement is $3\Delta A_{\mathrm{I,I}} + 4\Delta A_{\mathrm{I,II}}$. Compare this with the sequence $\textcircled{I}\textcircled{I}\textcircled{I}\textcircled{II}\textcircled{II}\textcircled{I}\textcircled{I}\textcircled{I}$ whose interaction energy is $4\Delta A_{\mathrm{I,I}} + \Delta A_{\mathrm{II,II}} + 2\Delta A_{\mathrm{I,II}}$. The energy difference between these two sequences is

$$\Delta A_{\mathrm{I,I}} + \Delta A_{\mathrm{II,II}} - 2\,\Delta A_{\mathrm{I,II}} \equiv \Delta_4 A_{\mathrm{I,II}}, \tag{8}$$

a relation well-known from the theory of mixtures. This rearrangement energy (8) is the same if there are molecules I all around those linear sequences, i.e., as if the arrangements are three-dimensional rather than one-dimensional as indicated.

This energy is negative definite which means that the second type of arrangement is energetically preferable to the first one. This negative definiteness of the rearrangement energy (8) is essentially due to the fact that the London energy $\Delta A_{I,II}$ between a pair of molecules I, II is proportional to the product $-\alpha_{\mathrm{I}}\alpha_{\mathrm{II}}$ of their polarizabilities as seen in (3) and (6). The rearrangement energy is then proportional to

$$-\alpha_{\mathrm{I}}^2 - \alpha_{\mathrm{II}}^2 + 2\alpha_{\mathrm{I}}\alpha_{\mathrm{II}} = -(\alpha_{\mathrm{I}} - \alpha_{\mathrm{II}})^2. \tag{9}$$

Turning our attention to the expression (9) we would like to mention that, because of (3) and (6), this expression becomes, in the classical and in the quantum limits, respectively,

$$\Delta_4 A_{\mathrm{I,II}} = -3R^{-6}kT(\alpha_{\mathrm{I}} - \alpha_{\mathrm{II}})^2 \leq 0, \tag{10}$$

$$\Delta_4 A_{\mathrm{I,II}} = -\tfrac{3}{4}R^{-6}\hbar\{(\alpha_{\mathrm{I}}\tilde{\omega}_{\mathrm{I}} - \alpha_{\mathrm{II}}\tilde{\omega}_{\mathrm{II}})^2 + \tilde{\omega}_{\mathrm{I}}\tilde{\omega}_{\mathrm{II}}(\alpha_{\mathrm{I}} - \alpha_{\mathrm{II}})^2\}/(\tilde{\omega}_{\mathrm{I}} + \tilde{\omega}_{\mathrm{II}}) \leq 0. \tag{11}$$

In using the oscillator representation, an actual molecule is composed of a sizable number of oscillators which are characterized by a distribution of polarizability over a broad region, mainly of the extreme ultraviolet frequency spectrum. Also, these oscillators have diversified anisotropy. A relation more general than (10) and (11) holds in such a situation and the inequality becomes approximately an equality only if the variety of parameters characterizing the frequency and directional distribution of oscillator polarizabilities are very similar in the two molecules I and II. This means that there is a multiparametric inequality, i.e., a multiparametric specificity for preferential association of identical or near identical molecules as nearest neighbors. The rearrangement energy then becomes

$$\Delta_4 A_{\mathrm{I,II}} = -\tfrac{1}{2}kT \sum_{S=-\infty}^{+\infty} \sum_{\mu,\nu=1}^{3}\sum [W_{S\mu\nu_{\mathrm{I}}} - W_{S\mu\nu_{\mathrm{II}}}]^2 \leq 0$$

which is a sum over S, i.e., the integral over the purely imaginary axis $\omega = iS2\pi kT/\hbar$. This was discussed in our previous papers (Bade, 1958; Yos, 1958; Jehle *et al.*, 1958). From known values of the London–van der Waals interactions (London, 1936) $\Delta A_{\mathrm{I,II}}$ can be estimated, and also an estimate can be made about the differential effect $\Delta_4 A_{\mathrm{I,II}}$ (Jehle, 1959, 1963; Bade, 1958; Yos, 1958). It is found to be much above the thermal noise level for a molecular side chain like an aromatic one. In the actually most interesting situations, sequences of pairwise identical side chains may associate and thus bring about a cumulative effect of many pair interaction terms, resulting in an interaction of higher energy and in a high degree of specificity.

The preferential association of identical or almost identical molecules as nearest neighbors is in evidence in several important biological phenomena. Muller (1922, 1937, 1941, 1947, 1950) pointed to such effects. The most important prototype of these kinds of phenomena is the tissue-specific cellular attachment of like cells to each other within an organism. On the molecular level the broad phenomenon of self-recognition is a point at issue [immunospecific interactions (Haurowitz, 1950, 1963), enzyme complementation]; and the problem of replication of nucleic acids seems to depend on that mechanism in part, too.

ACKNOWLEDGMENTS

This work was supported by USPHS grant #CA-04989-05 and GM-12054 and by the George Washington University Committee on Research. The author is greatly indebted to discussions and suggestions made to him at the Stanford Symposium on Quantum Aspects of Polypeptides and Polynucleotides, at the University of Uppsala, at the Ravello Symposium on Electronic Aspects of Biochemistry, and at the Squaw Valley Summer School on Molecular Biophysics, and in particular to the hosts, Dr. M. Weissbluth, Dr. P. O. Löwdin, and Dr. B. Pullman. My special thanks go to my students here at George Washington University who shared in the work, in particular Mr. W. C. Parke, Mr. M. L. Ingerman, and Miss A. Salyers.

REFERENCES

ABRIKOSOV, A. A., GORKOV, L. P., and DZIALOSHINSKII, I. E. (1959). *Zh. Eksperim. i Teor. Fiz.* **36**, 900; (1957) **33**, 799.

ABRIKSOV, A. A., GORKOV, L. P., and DZIALOSHINSKÜ, I. E. (1963). "Methods of Quantum Field Theory in Statistical Physics" (R. A. Silverman, ed.). Prentice-Hall, Englewood Cliffs, New Jersey.

BADE, W. L. (1957). *Proc. Nat. Acad. Sci. U.S.* **43**, 341.

BADE, W. L. (1957a). *J. Chem. Phys.* **27**, 1280.

BADE, W. L. (1958). *Phys. Rev.* **110**, 793.

BADE, W. L. (1958a). *J. Chem. Phys.* **28**, 282.

BADE, W. L., and KIRKWOOD, J. G. (1957). *J. Chem. Phys.* **27**, 1284.

BECKER, R. (1955). "Theorie der Wärme." Springer, Berlin (rev. ed., 1960).

BRIEGLEB, G. (1961). "Elektronen-Donator-Acceptor-Komplexe." Springer, Berlin.

CALLEN, H. B., and WELTON, T. A. (1951). *Phys. Rev.* **83**, 34.

CASIMIR, H. B. C. (1948). *Konink. Ned. Akad. Wetenschap. Proc., Ser. B*, **60**, 793.

DE BOER, J. H., and HELLER, G. (1937). *Physica* **4**, 1045.

DEBYE, P. (1920). *Physik. Z.* **21**, 178; **22**, 302.

DEBYE, P., and HÜCKEL, E. (1923). *Physik. Z.* **24**, 185.

DERJAGUIN, B. U., ABRIKOSOVA, I. I., and LIFSHITZ, E. M. (1956). *Quart. Rev. Chem. Soc. (London)* **10**, 295.

DERJAGUIN, B. U., ABRIKSOVA, I. I., and LIFSHITZ, E. M. (1960). *Sci. Am.* **203**, 47.

DOTY, P. (1959). *Rev. Mod. Phys.* **31**, 107.

DOTY, P., MARMUR, J., EIGNER, J., and SCHILDKRAUT, C. (1960). *Proc. Natl. Acad. Sci. U.S.* **46**, 461.

DZIALOSHINSKII, I. E. (1957). *Soviet Phys. JETP (Engl. Transl.)* **3**, 977.

DZIALOSHINSKII, I. E., and PITAEVSKII, L. P. (1959). *Soviet Phys. JETP (Engl. Transl.)* **9**, 912.

DZIALOSHINSKII, I. E , LIFSHITZ, E. M., and PITAEVSKII, L. P. (1959). *Zh. Eksperim. i Teor. Fiz.* **37**, 229.

EIGEN, M., and DE MAYER, L. (1958). *Proc. Roy. Soc.* **A247**, 505.

EISENSCHITZ, R., and LONDON, F. (1930). *Z. Physik* **60**, 491; **63**, 245.

EYRING, H., WALTER, J., and KIMBALL, G. E. (1944). "Quantum Chemistry." Wiley, New York.

FEYNMAN, R. P. (1939). *Phys. Rev.* **56**, 340.

FÖRSTER, T. (1951). "Fluoreszenz organischer Verbindungen." V. & R., Goettingen.

FÖRSTER, T. (1959). *Discussions Faraday Soc.* **27**, 7.

FÖRSTER, T. (1963). *Rev. Mod. Phys.* **35**, 572.

FRANK, H. S. (1958). *Proc. Roy. Soc.* **A247**, 481.

FRANK, H. S., and WEN, W. T. (1957). *Discussions Faraday Soc.* **24**, 133.

GINSBURG, V. L., and PITAEVSKII, L. P. (1958). *Soviet Phys.* (*Engl. Transl.*) **34**, 1240.

HAMAKER, H. C. (1937). *Physica* **4**, 1058.

HAMAKER, H. C. (1938). *Chem. Weekblad* **35**, 47.

HAMAKER, H. C. (1948). *In* "Theory of the Stability of Lyophobic Colloids" (E. J. W. Verwey and J. T. G. Overbeek, eds.), p.161. Elsevier, Amsterdam.

HAUROWITZ, F. (1950). "The Chemistry and Biology of Proteins." Academic Press, New York (2nd ed., 1963: "The Chemistry and Function of Proteins").

HEITLER, W. (1945). "Elementary Wave Mechanics." Oxford Univ. Press, London and New York.

HELLMAN, H. (1937). "Quantenchemie." Deuticke, Vienna.

HIRSCHFELDER, J. O. (1941). *J. Chem. Phys.* **9**, 645.

HIRSCHFELDER, J. O. (1963). *Rev. Mod. Phys.* **35**, 616.

HIRSCHFELDER, J. O. (1964). *Univ. Wisconsin* (*Madison, Wis.*) *Theoret. Chem. Rept.* No. WIS-TCI-66.

HIRSCHFELDER, J. O. (1965). "Molecular Biophysics."

HIRSCHFELDER, J. O., and ELIASON, M. A. (1957). *Ann. N.Y. Acad. Sci.* **67**, 451.

HIRSCHFELDER, J. O., CURTISS, C. F., and BIRD, R. B. (1954). "Molecular Theory of Gases and Liquids." Wiley, New York.

JEHLE, H., (1959). *Proc. Natl. Acad. Sci. U.S.* **45**, 1360.

JEHLE, H. (1963). *Proc. Natl. Acad. Sci. U.S.* **50**, 516 and 738.

JEHLE, H. (1965). *Proc. Nat. Acad. Sci. U.S.* **53**. 1451

JEHLE, H., YOS, J. M., and BADE, W. L. (1958). *Phys. Rev.* **110**, 793.

JEHLE, H., PARKE, W. C., and SALYERS, A. S. (1964a). *Biophysika* **9**, 401.

JEHLE, H., PARKE, W. C., and SALYERS, A. S. (1964b). "Electronic Aspects of Biochemistry," p. 313. Academic Press, New York.

JEHLE, H., PARKE, W. C., SHIRVEN, R. M., and AEIN, D. K. (1964c). *Biopolymer Symp.* No. 1, p. 209. Wiley (Interscience), New York.

KASHA, M. (1964). "Studies in Quantum Chemistry." Inst. of Molecular Biophys., Tallahassee, Florida.

KAUZMANN, W. (1954). *In* "Mechanism of Enzyme Action," Symposium (W. D. McElroy and B. Glass, eds.), p.70, Johns Hopkins Press, Baltimore, Maryland.

KAUZMANN, W. (1957). "Quantum Chemistry." Academic Press, New York.

KEESOM, W. H. (1912). *Commun. Phys. Lab. Univ. Leiden* Suppl. 24–26.

KESTNER, N. R., and SINANOĞLU, O. (1963). *J. Chem. Phys.* **38**, 1730.

KIRKWOOD, J. G., and SHUMAKER, J. B. (1952). *Proc. Natl. Acad. Sci. U.S.* **38**, 863.

KLOTZ, I. M. (1962). *In* "Horizons in Biochemistry," Albert Szent-Györgyi Dedicatory Volume (M. Kasha and B. Pullman, eds.), p.523. Academic Press, New York.

KRAMERS, H. A., and KRONIG, R. de L. (1928). *Z. Physik* **48**, 174.

KRAMERS, H. A., and KRONIG, R. DE L. (1956). *In* "Collected Scientific Papers of Kramers," p.382.

KRUYT, H. R., and OVERBEEK, J. T. G. (1964). "Introduction to Physical Chemistry." Holt, New York.

LANDAU, L. D. (1958). *Zh. Eksperim. i Teor. Fiz.* **34**, 262.

LANDAU, L. D., and LIFSHITZ, E. M. (1958). "Statistical Physics." Addison-Wesley, Reading, Massachusetts.

LANDAU, L. D., and LIFSHITZ, E. M. (1960). "Electrodynamics of Continuous Media." Addison-Wesley, Reading, Massachusetts.

LIFSHITZ, E. M. (1955). *Zh. Eksperim. i Teor. Fiz.* **29**, 94.

LIFSHITZ, E. M., DZIALOSHINSKII, I. E., and PITAEVSKII, L. P. (1961). *Advan. Phys.* **10**, 165.

LINDER, B. (1960). *J. Chem. Phys.* **33**, 688.

LINDER, B. (1962). *J. Chem. Phys.* **37**, 963.

LINDER, B. (1964). *J. Chem. Phys.* **40**, 622 and 2003.

LINDERBERG, J. (1962). *Phys. Letters* **1**, 272.

LINDERBERG, J., and SHULL, H. (1960). *J. Mol. Spectry.* **4**, 30; **5**, 1.

LONGUET-HIGGINS, H. C. (1956). *Proc. Roy. Soc.* **A235**, 537.

LÖWDIN, P.-O. (1955). *Phys. Rev.* **97**, 1474.

LÖWDIN, P.-O. (1956). *Texas J. Sci.*, **8**, 163.

LÖWDIN, P.-O. (1957). *J. Phys. Chem.* **61**, 55.

LÖWDIN, P.-O. (1960). *Ann. Rev. Phys. Chem.* **11**, 107.

LÖWDIN, P.-O. (1962). *Rev. Mod. Phys.* **34**, 80 and 520.

LÖWDIN, P.-O. (1963). *Rev. Mod. Phys.* **35**, 702.

LONDON, F. (1936). *Trans. Faraday Soc.* **33**, 8.

LONDON, F. (1942). *J. Chem. Phys.* **46**, 305.

MCLACHLAN, A. D. (1963). *Proc. Roy. Soc.* **A271**, 387.

MARGENAU, H. (1939). *Rev. Mod. Phys.* **11**, 1.

MATSUBARA, T. (1955). *Progr. Theoret. Phys.* (*Kyoto*) **14**, 351.

MULLER, H. J. (1922). *Am. Naturalist* **56**, 32.

MULLER, H. J. (1937). *Sci. Monthly* **44**, 210.

MULLER, H. J. (1941). *Cold Spring Harbor Symp. Quant. Biol.* **9**, 290.

MULLER, H. J. (1947). *Proc. Roy. Soc.* **B134**, 1.

MULLER, H. J. (1950). *In* "Genetics in the 20th Century" (L. C. Dunn, ed.), p.77. Macmillan, New York.

MULLIKEN, R. S. (1950). *J. Am. Chem. Soc.* **72**, 600.

MULLIKEN, R. S. (1951). *J. Chem. Phys.* **19**, 514.

MULLIKEN, R. S. (1952). *J. Am. Chem. Soc.* **74**, 811.

MULLIKEN, R. S., and PERSON, W. B. (1962). *Ann. Rev. Phys. Chem.* **13**, 107.

NEMETHY, G., and SCHERAGA, H. A. (1963). *Biopolymers* **1**, 43.

NOZIERES, P., and PINES, D. (1958). *Phys. Rev.* **109**, 762; *Nuovo Cimento* **9**, 470.

NYQUIST, H. (1928). *Phys. Rev.* **32**, 110.

ONCLEY, J. L., and SCHEIDER, W. (1964). Personal communication.

ONSAGER, L. (1933). *Chem. Rev.* **13**, 73.

ONSAGER, L. (1936). *J. Am. Chem. Soc.* **58**, 1486.

OVERBEEK, J. T. G. (1948). *In* "Theory of Lyophobic Colloids" (E. J. W. Verwey and J. T. G. Overbeek, eds.), p.161. Elsevier, Amsterdam.

OVERBEEK, J. T. G. (1952). *In* "Colloid Science" (H. R. Kruyt, ed.), Vol. 1, p.276. Elsevier, Amsterdam.

PAULING, L. (1940). *J. Am. Chem. Soc.* **62**, 2643.

PAULING, L. (1948). *Nature* **161**, 707.

PAULING, L. (1955). "Aspects of Synthesis and Order in Growth," p.3. Princeton Univ. Press, Princeton, New Jersey.

Pauling, L. (1957). "Probability of Errors in the Process of Synthesis of Protein Molecules," Stoll Festschrift, p.597. Birkhäuser, Basel.

Pauling, L. (1960). "The Nature of the Chemical Bond," 3rd ed. Cornell Univ. Press, Ithaca, New York.

Pauling, L. (1960b). *In* "Aspects of the Origin of Life" (M. Florkin, ed.), p.132. Macmillan (Pergamon), New York.

Pauling, L., and Delbrück, M. (1940). *Science* **92**, 77.

Pauling, L., and Hayward, R. (1964). "The Architecture of Molecules." Freeman, San Francisco, California.

Pauling, L., and Itano, H. A. (1957). "Molecular Structure and Biological Specificity." Am. Inst. Biol. Sci., Washington, D.C.

Pauling, L., and Wilson, E. B., Jr. (1935). "Introduction to Quantum Mechanics." McGraw-Hill, New York.

Pauling, L., Campbell, D. H., and Pressman, D. (1943). *Physiol. Rev.* **23**, 203.

Pimentel, G. C., and McClellan, A. L. (1960). "Hydrogen Bond." Freeman, San Francisco, California.

Pines, D. (1955). *Solid State Phys.* **1**, 368.

Platzman, R. L. (1962). *Vortex* **28**, 372.

Pullman, A., and Pullman, B. (1963). "Quantum Biochemistry." Wiley (Interscience), New York.

Rabinowitch, E., and Epstein, L. F. (1941). *J. Am. Chem. Soc.* **63**, 69.

Rhodes, W. (1962). *J. Chem. Phys.* **37**, 2433.

Rostow, S. M. (1963). "Theory of Electromagnetic Fluctuations" [in Russian]. Academy of Sciences, Moscow.

Salem, L. (1964). *In* "Electronic Aspects of Biochemistry," Symposium (B. Pullman, ed.), p.293. Academic Press, New York.

Scatchard, G. (1931). *Chem. Rev.* **8**, 321.

Scatchard, G. (1937). *Trans. Faraday Soc.* **33**, 160.

Scatchard, G. (1952). *Am. Scientist* **40**, 61.

Sinanoğlu, O., Abdulner, S., and Kestner, N. R. (1964). *In* "Electronic Aspects of Biochemistry," Symposium (B. Pullman, ed.), p.301. Academic Press, New York.

Slater, J. C. (1963). "Quantum Theory of Molecules and Solids," Vol. 1. McGraw-Hill, New York.

Tanford, C. (1961). "Physical Chemistry of Macromolecules." Wiley, New York.

Tinoco, I. (1960). *J. Am. Chem. Soc.* **82**, 4785.

Tinoco, I. (1960a). *J. Chem. Phys.* **33**, 1332.

Tinoco, I., and DeVoe, H. (1962). *J. Mol. Biol.* **4**, 518.

Tomonaga, S. (1950). *Progr. Theoret. Phys.* (*Japan*) **5**, 554.

Tomonaga, S. (1955). *Progr. Theoret. Phys.* (*Japan*) **13**, 467.

Van Vleck, J. H., and Sherman, A. (1934). *Rev. Mod. Phys.* **7**, 167.

Vinograd, J. R. (1935). *In* "Thixotropy" (H. Freundlich, ed.), p.267. Herman, Paris.

Wang, S. C. (1927). *Physik. Z.* **28**, 663.

Weissbluth, M. (1961). "Optical Properties of Polymers," Biophys. Lab. Rept. No. 19. Stanford Univ. Press, Stanford, California.

Yos, J. M. (1958). *Phys. Rev.* **110**, 800.

Yos, J. M., Bade, W. L., and Jehle, H. (1957). *Proc. Nat. Acad. Sci. U.S.* **43**, 341 and 847.

Wigner, E. (1934). *Phys. Rev.* **46**, 1002.

Wigner, E. (1938). *Trans. Faraday Soc.* **34**, 678.

Quantum Genetics and the Aperiodic Solid

Some Aspects on the Biological Problems of Heredity, Mutations, Aging, and Tumors in View of the Quantum Theory of the DNA Molecule*

PER-OLOV LÖWDIN

Department of Quantum Chemistry
Uppsala University
Uppsala, Sweden
and
Quantum Theory Project
University of Florida
Gainesville, Florida

---

* Sponsored in part by the King Gustaf VI Adolf's 70-Years Fund for Swedish Culture, Knut and Alice Wallenberg's Foundation, and in part by the Cancer Institute of National Institutes of Health, under Contract CA 06850-01 with Uppsala University.

Recent development in biology has shown that deoxyribonucleic acid (DNA) is the hereditary substance which carries the genetic message in the cell. After a brief historical review of the development, the structure of DNA is discussed in greater detail.

According to Watson and Crick's stereomodel (Watson and Crick, 1953; Crick and Watson, 1954), the DNA molecule is a double helix with the two sugar-phosphate chains held together by a sequence of pairs of nucleotide bases (A, G, T, C) joined by hydrogen bonds. Depending on the structure of these bonds, the pairing is specific (A—T, G—C), and, in the DNA duplication, each half of the helix serves as a template in forming its own complement. It is believed that the genetic information is contained in the sequence of base pairs which is preserved through this arrangement. Watson and Crick have also pointed out that the bases may occur in rare tautomeric forms A*, T*, G*, and C*, which have another pairing scheme A*—C, G*—T, etc., depending on the changed possibilities for the formation of hydrogen bonds. Incorporation of these tautomeric forms may then lead to errors in the genetic code leading to mutations.

Through a series of experiments, one has proven that the DNA replication is semiconservative and that the two strands in the original double helix separate and go into different daughter helices with the help of a unwinding-winding mechanism described by the Y-model; a survey of the literature is given. The template idea is hence firmly supported. Recent measurements of the tautomeric equilibrium constants for some of the bases indicate, however, that the tautomeric forms are by no means as rare as expected and that one could anticipate numerous errors to occur at every DNA replication. It is here suggested that this difficulty may be removed by assuming the existence of a temporary "replication plane" at the branching point of DNA in the Y-model which, after separation of the two strands by a rotation of one of them 180° and addition of two new nucleotides, will contain four nitrogen bases. If the plane contains two amino groups and two keto groups around the axis of the helix, there is a possibility for a formation of temporary hydrogen bonds which make such a plane energetically favorable in comparison to the planes containing three amino groups or three keto groups. The arrangement will hence essentially prevent the incorporation of single tautomeric bases from the environment but not tautomeric pairs. The probability for the simultaneous occurrence of a tautomeric pair is, however, very small and apparently in good agreement with the order of magnitudes of the mutation rates.

Each hydrogen bond in DNA consists of a proton shared between two electron lone pairs, and the genetic code is essentially a "proton code." The probability for proton transfer in the hydrogen bonds of DNA is further discussed. By using the available charge orders for the π electrons of the base pairs, it is shown that the double-well potentials acting on the protons are highly asymmetric. At normal temperature ($T = 310°\text{K}$), there is practically no proton transfer above the barrier, explaining the enormous stability of the genetic code. According to quantum mechanics, however, a proton is not a classical particle but a "wave packet" which may penetrate a potential barrier by means of the "tunnel effect." Depending on this proton tunneling, there is hence a very small but with time increasing probability that the normal base pairs A—T and G—C may spontaneously go over into the tautomeric pairs A*—T* and G*—C* through a "proton exchange" along the hydrogen bonds. Since the tautomeric bases have another pairing pattern, the proton exchange leads inevitably to errors in the genetic base sequence in the next duplication.

In neutral base pairs, the proton tunneling is a very slow process but, if one of the bases becomes charged, it may turn into a comparatively fast process. This means that electron or proton transfer in DNA caused by radiation or chemical reactions may greatly speed up the loss of the genetic code, and certain "induced phenomena" are discussed from this point of view.

The various possibilities suggested for the transcription of the genetic code through the formation of messenger RNA are studied, and the present

status of the coding problem in protein synthesis is briefly reviewed. In conclusion, the biological implications of the proton tunneling with respect to the phenomena of aging, spontaneous and induced mutations, and occurrence of tumors are discussed.

I. Introduction

The cell may be considered as the fundamental biological unit. The cell consists of a nucleus surrounded by cytoplasm, in which further various bodies may occur. The nucleus is responsible for the most important properties of the living cell. The fundamental properties of the nucleus are determined by the chromosomes containing the genes, and these consist essentially of giant organic molecules known as deoxyribonucleic acid (DNA), which are the essential carriers of the genetic information characteristic for the species. Many biological and medical problems are directly related to the properties of these molecules and, since they govern the cell duplication and the protein synthesis, they are of essential importance in the problem of normal growth and aging as well as in the development of tumors.

After many years of experimental research, the structure of DNA is now well-known. According to Wilkins–Watson–Crick's stereomodel, DNA consists of a double helix, where the sides or strands are sugar-phosphate chains joined by pairs of nucleotide bases. Each base pair consists of a purine base and a pyrimidine base held together by two hydrogen bonds. The DNA problem is greatly simplified by the fact that there are only four nucleotide bases involved: adenine (A), thymine (T), guanine (G), and cytosine (C). The pairing of the bases is further specific in the combinations A—T, G—C, so that each base has a specific complementary base.

It is today believed that the genetic information is contained in the sequence of the nucleotide bases attached to one strand of the DNA molecule, i.e., in a four-letter message of the form A G T C A T T G C A .... The other half contains the complementary sequence, i.e., T C A G T A A C G T ..., and the double helix consists of two parts which may be compared with a photographic negative and positive. In the mechanism of cell duplication, the hydrogen bonds in the double helix are released, the two DNA strands become at least partly free, and each one produces by means of enzymes and available material its own complement, so that two complete DNA molecules with the same base sequence as the original one are produced.

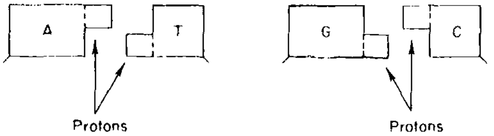


Fig. 1. Schematic form of the nucleotide bases occurring in DNA.

The complementary property of the bases depends on certain steric factors which are schematically described in Fig. 1. There are two "long" bases of purine character, A and G, and two "short" bases of pyrimidine type, T and C. Each base offers a surface which contains a proton in a characteristic position. The replication of a base pair is indicated in Fig. 2.

The complementary nature of the DNA molecule explains the stability of the genetic message and how it is propagated at a cell duplication. It also tells that if there is an error in the base sequence and the corresponding genetic message, such an error may undergo "biological amplification" by the factors 2, 4, 8, 16, 32, ... under the ordinary cell duplication. If such an error occurs in the first cell of an individual, there may be a *mutation* whereas, if it occurs at a later stage as a somatic mutation, there may develop a "new individual" within the individual in the form of a tumor. It is well-known that such changes in the genetic information may be caused by chemicals or by physical means, e.g., radiation. The occurrence of spontaneous mutations and tumors in nature implies further that *spontaneous errors* must sometimes occur in the genetic message, i.e., in the DNA molecule, but the nature of these errors is so far not known.

In this paper, the stability of the DNA molecule and the occurrence of spontaneous and induced errors will be studied. According to the Wilkins–Watson–Crick model, the genetic message is essentially contained

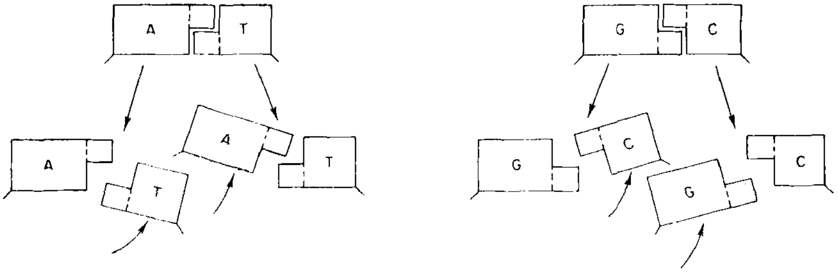


Fig. 2. Schematic diagram of the replication of a base pair in DNA. Curved arrows represent nucleotides from the environment.

in the "proton code" occurring in the hydrogen bonds in the middle of the double helix. The nucleotide bases are planar conjugated systems in which the protons are attached to electron lone pairs. In DNA, proton errors may occur both during replication and transcription and during the biologically inactive period. In all these connections, certain isomeric forms of the bases play a fundamental role. The tautomeric forms A*, G*, T*, and C* are isomers obtained by moving a proton from one electron lone pair to another, and the schematic forms are outlined in Fig. 3.

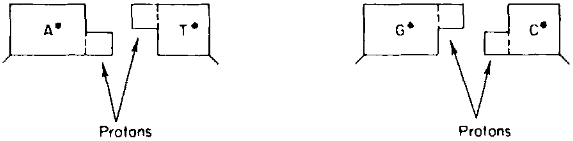


Fig. 3. Schematic form of the nucleotide bases in their rare tautomeric forms.

Depending on the changed possibilities for the formation of hydrogen bonds, they have another pairing pattern than normal, namely A*—C, G*—T, T*—G, and C*—A. Watson and Crick suggested that the occurrence of these tautomeric forms may be responsible for the phenomenon of mutation.

In DNA replication, the new bases used in the synthesis of the daughter molecules are essentially in normal form and, since mutations are rare events, it was assumed that the tautomeric forms were exceedingly rare. Recent measurements of the tautomeric equilibrium constants for the pyrimidine bases (Katrizky and Waring, 1962) have shown, however, that somewhat less than 1‰ of these bases may occur in the tautomeric forms. This means that, since a DNA molecule usually consists of more than a million base pairs, there will be several hundred tautomeric errors per replication. Experience tells us that the replication mechanism is remarkably exact, and that such a large number of incorporation errors in some way is avoided.

In order to solve this problem, it is necessary to consider the Y-model for the DNA replication in detail and particularly the branching point and building area, in which four bases come close to each other in a "replication plane." Chargaff has many times pointed out the importance of the pairing between amino and keto groups, and this idea can now be fully utilized. The bases A and C (as well as their tautomeric forms) are 6-amino bases, whereas G and T and their tautomers are 6-keto bases. It will be shown that a replication plane containing two keto and two amino bases in a parallelogram, depending on temporary hydrogen bonding, is energetically much more favorable than a replication plane containing three amino or three keto groups. This simple mechanism prevents the incorporation of a single tautomeric base. A tautomeric "pair" may still be incorporated, but the probability for the occurrence of such a pair is much lower and in reasonable agreement with the low mutation rates per base pair and generation estimated from experience (10^{-8} to 10^{-11}, Freese, 1962).

Of interest also is the change of the genetic code in DNA during the biologically inactive periods associated with physical processes. Each hydrogen bond is a proton shared between two electron lone pairs. The potential acting on the proton is a double-well potential with two minima, and the large asymmetry between the minima gives the code the necessary stability. For most chemical purposes, it is sufficient to consider the protons as positive point charges. Modern physics tells us, however, that these protons are actually "wave packets" obeying the laws of quantum theory. Such a wave packet may leak through a potential barrier which is a complete obstacle for a classical particle, and this phenomenon is known as the quantum-mechanical "tunnel effect." Even if the genetic code immediately after a replication is "pure," there starts a temperature- and time-dependent process associated with the proton tunneling which corresponds to a slow but gradual loss of the genetic code. In order to keep the gross electric neutrality, two protons are usually tunneling almost simultaneously in opposite directions leading to the formation of a tautomeric pair, as indicated in Fig. 4. The corresponding replication scheme has the form shown in Fig. 5, showing that the proton exchange leads to the mutation A—T→G—C, and vice versa. The proton tunneling is a slow spontaneous process which is characteristic for a base pair and its electrostatic environment, and it depends hence also on the neighboring base pairs and on the properties of the molecule as a whole. In molecules which are damaged by radiation or certain chemical processes, the proton tunneling may turn into a fast process of importance in understanding induced phenomena.

Fig. 4. Schematic diagram of formation of a tautomeric base pair through proton tunneling.

It is well-known that the genetic code may be influenced, e.g., by

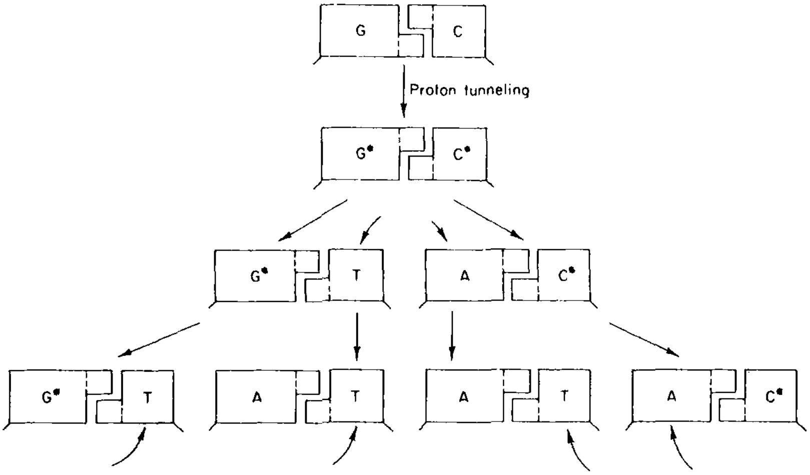


Fig. 5. Replication of a base pair in DNA after proton tunneling.

electromagnetic radiation, by electron and proton bombardment, and by certain chemicals known as mutagens and carcinogens. Certain mutagens, as nitrous acid, may cause a rather large change in the chemical structure of the bases and hence in the proton code, but otherwise the obvious chemical damage is usually very small. It seems as if the main effect would simply be a removal or addition of a single charge through ionizing radiation or donor-acceptor reactions. If one of the bases in a pair becomes charged in comparison to the other, one of the two double-well potentials in the hydrogen bonds loses suddenly part of its asymmetry, and the probability for a single proton transfer through a tunnel effect is highly increased. In this case, one has transitions of the type indicated in Fig. 6 and it should be observed that the resulting "ions" A^-, T^+, G^-, C^+, etc., do not combine with the normal bases; the code is not only lost, but it is replaced by a repulsive element. If these ions occur in connection with a DNA replication, it may happen that they do not find complementary

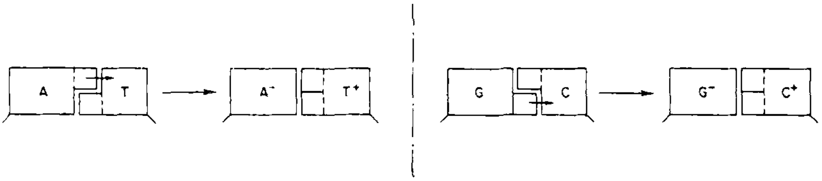


Fig. 6. Change of the genetic code due to single proton tunneling in a charged based pair.

bases, which means that there will be a "deletion" on the other strand of the new molecules, see Fig. 7. In many cases, the separated ions will, of course, catch or lose protons, but the genetic code may still be disturbed. Since neighboring base pairs have partly overlapping π electron clouds, there is a helical pathway for electrons along the axis, i.e., a certain semiconductivity, and this implies that, if a donor-acceptor reaction occurs at one end of DNA, the electron involved may still travel a considerable distance along the molecule causing damage to the genetic code before it is finally trapped.

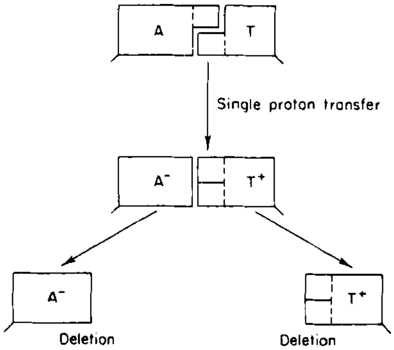


Fig. 7. Replication of a base pair after single proton tunneling.

The biological effect of spontaneous proton tunneling is highly different for different species. In an organism consisting of a single cell, the tunneling contributes to the occurrence of spontaneous mutations. In a multicell organism, there is a great deal of differentiation between the cells; depending on their functions, certain cells divide frequently, whereas other cells (as the nerve cells and the brain cells of higher organisms) stop their replication at a certain stage of life. If the cells divide frequently, a mutation caused by proton tunneling or an incorporation error will usually cause little harm, since the corresponding cell will either die directly or lose in competition with its normal neighbors. Only during exceptional circumstances will the new cell have obtained such a code that it defeats its normal neighbors, wins the struggle for existence, starts to replicate, and forms the nucleus of a *tumor*. Today it is rather generally believed that somatic mutations are the original cause of cancer, but this problem has also many aspects dealing with the organization of the organism as a whole as a superstructure of highly differentiated cells.

In cells which have stopped their replication in an early stage of life, the proton tunneling in DNA leads to a gradual loss of the genetic code which is different for different molecules. The code is still used in transcription in forming the messenger RNA governing the protein synthesis but, if part of the code is lost, the cell may lose its ability to produce the necessary proteins and particularly the enzymes. Such a cell may die and, since it is not replaced, the entire organism suffers a certain damage which may be one of the primary causes for the phenomenon of *aging*.

If this interpretation is correct, the biological phenomena of aging and occurrence of spontaneous mutation and tumors would to a large extent depend essentially on the quantum-mechanical "tunnel effect"—a conclusion which may at first sight seem rather startling. However, the idea agrees very well with the opinion previously expressed by Delbrück and by Schrödinger that a mutation is associated with a "quantum jump" in the molecule carrying the genetic information. Even if the model is certainly highly oversimplified, it adds new aspects to the problem of the nature of life which may be worthwhile investigating in somewhat greater detail.

This introduction shows that the problem we are going to study in this paper is on the borderline between several different fields: mathematics, physics, chemistry, biochemistry, biology, and medicine. It is clear that no one can be a specialist in all these areas, and many subjects can be covered only in a very rough and certainly inadequate way. Some people claim that only chemists can speak to chemists, biologists to biologists, etc., and that nonspecialists should keep quiet since they will anyway only stir up confusion. Today the natural sciences are so highly overlapping, however, that I feel that, in spite of all difficulties, it is greatly important to try to bridge the gap between the classically established fields and create a single unified area: science.

The author would like to apologize to the many specialists for the obvious weaknesses of his treatment. At the same time, he hopes that this little survey may stimulate some scientists to look more carefully into the many interesting problems associated with the borderline field of the nature and properties of the genetic code.

II. Historical Development of the Theory of the Hereditary Mechanism

The present knowledge of the nature of the genetic information in all living materials is the result of the collaboration of many different sciences, and we will start with a brief review of the historical development.

A. Cell Theory and Genetics

Optics and the art of lens grinding should perhaps be considered as the first contributor to the modern development, since it led to the construction of the microscope. By means of such an instrument, the cell was discovered by Robert Hooke in 1665. In the beginning of the 19th century, the technique was developed far enough to permit a study of the details of the cell. The cell nucleus was discovered in 1831 by R. Brown, and the cell division was first described by H. v. Mohl in 1835. The general cell theory for plants was formulated by M. J. Schleiden in 1838, and the theory was extended to all living organisms by T. Schwann in 1839. Since this date, the cell theory has had an enormous impact on the entire biology as being one of the most important generalizations used in describing life processes.

At about the same time as Charles R. Darwin (1859) published his *Origin of Species*, the Abbot Gregor Mendel (1865) made experiments on the hereditary properties of some varieties of garden pea in his cloister in Brünn. In contrast to Darwin's theory of evolution which aroused an enormous interest even far outside biology, Mendel's work published in 1865 remained practically unnoticed until the turn of the century.

In the meantime, there had been an important development in the study of the cells. Thanks to refined dyeing processes in microscopic technique, one had been able to investigate the cell nucleus in greater detail and discovered the "chromosomes." The specificity of the chromosomes and their behavior under ordinary cell division (mitosis) and under reductive division (meiosis) was first described by É. van Beneden, C. Rabl, T. Boveri, and E. Strasburger in 1883–1888.

In 1900 Mendel's work on the law of inheritance was rediscovered independently by H. de Vries, C. Correns, and E. Tschermak, and was made the basis for the new science of genetics. The three papers by these authors appear remarkably enough in the same volume of *Berichte der Deutschen Botanischen Gesellschaft.* A big step forward could then be taken by combining genetics with cytology and, assuming that the chromosomes are the essential carriers of the genetic information, one could derive a picture of the hereditary mechanism in complete agreement with Mendel's laws.

In 1902, de Vries discovered the mutations. Even in the most homogeneous biological population, there are always small continuous variations of random character and, in his theory of evolution, Darwin had assumed that these accidental variations would provide sufficient material for the natural selection to work. It turned out, however, that these small continuous variations are not inherited. Instead de Vries found that, even in carefully pure-bred populations, there may be a few offsprings—say one in ten thousand—showing a small but discontinuous change which is then inherited. These "mutants" provide, of course, a still better material for Darwin's natural selection, but the nature of the mutation itself was very difficult to understand. One of the fundamental problems in genetics was hence to explain both the immense stability of the hereditary substance in the chromosomes over thousands of years as well as the occurrence of the discontinuous changes leading to the mutations.

After the discovery of the Mendelian laws, genetics developed rapidly and became soon a science of fundamental importance in biology. Here only the outstanding work by Th. H. Morgan and his school on the genetics of the fruit fly (*Drosophila*) will be mentioned. Assuming that the laws of heredity have a general character, the geneticists have during the last decades turned their interest to still smaller and more convenient subjects of investigation, such as fungi, bacteria, and viruses. Today the bacterium *Escherichia coli* and the bacteriophage T2 belong to the most thoroughly studied and well-known biological species. Moreover, one can observe a strong tendency of genetics to move over into a molecular and submolecular level in its attempts to explain the various phenomena and make new discoveries and explorations. For a survey, we will refer to Strauss (1960).

B. Quantum Theory of Matter

At this point, it may be worthwhile to review briefly some developments in modern physics. At the end of the 19th century, the classical physics appeared to be quite successful in all its applications except for a few minor discrepancies, for instance in the theory of "black-body radiation," where it seemed impossible to bring the results obtained for the long-wavelength part in agreement with those for the short-wavelength part. These difficulties were removed in one stroke by Max Planck in 1900 by the assumption that the energy of an harmonic electric oscillator emitting radiation could only be an integer multiple of the "energy quantum" $E = h\nu$, where ν is the frequency of the oscillator and h is the so-called Planck's constant or quantum of action: $h = 6.625 \times 10^{-27}$ ergsec. Hardly could Planck anticipate that this "quantum hypothesis" would soon change the fundaments of all of physics and a large part of chemistry. In 1905 Albert Einstein applied the new hypothesis to the electromagnetic waves and, by assuming that the radiation occurs quantized in the form of "wave packets," he could successfully explain the photoelectric effect.

By applying the quantum postulate to Rutherford's atom model, Niels

Bohr derived in 1913 his famous theory for the hydrogen atom consisting of an electron circling around a proton. Later the orbits were extended to be ellipses characterized by three quantum numbers (n, l, m). Bohr assumed further that the atom could exist only in certain "stationary states," and that the transitions between these states occurred as discontinuous "quantum jumps" connected with radiation. At this time, no one anticipated any connection between quantum theory and the basic laws of genetics, particularly since the physicists were fully busy extending the new approach to the many-electron atoms.

The Bohr–Rutherford atom model has been compared with a solar system in miniature, where the electrons move in elliptical orbits around a positively charged atomic nucleus of extremely small dimensions. The number of positive fundamental charges in the atomic nucleus is called the atomic number Z and equals the number of electrons in the neutral atom; it can be experimentally measured by means of Moseley's law for the K line in the X-ray spectra of the atoms formulated in 1913.

The rare gases He, Ne, A, Kr, Xe, Rn, ... are elements with the atomic numbers $Z = 2, 10, 18, 36, 54, 86, \ldots$, respectively, and, since all these elements are chemically inactive, one can assume that the corresponding electron configurations are particularly stable; they are characterized as "rare gas shells." In 1916, G. N. Lewis pointed out that the electrons in an atom are conveniently divided into two groups: the rare gas shell and the valence electrons, of which only the latter are chemically active. By considering the valence electrons in a molecule, Lewis could also show that each covalent chemical bond is associated with an electron pair shared between the two atoms involved. However, any deeper explanation of the stability of such a bond was still lacking and would come only through the further development of quantum theory.

In 1924, Louis de Broglie reversed Einstein's wave corpuscle parallelism of 1905 by assuming that each fundamental particle (electron, proton, etc.) is associated with a wave packet and, in this way, he could give a deeper explanation of Bohr's model for the hydrogen atom. However, the full content of the new approach was first realized by Erwin Schrödinger in 1925 who postulated the existence of a certain wave equation for the fundamental particles involved; with this work physics took the step over to wave mechanics or *modern quantum theory*. It is remarkable that the new theory was at the same time developed independently by two other authors: by W. Heisenberg in terms of matrices, and by P. A. M. Dirac in terms of q numbers.

Schrödinger could show that the existence of the stationary states of atoms and molecules was an immediate consequence of the theory, and that these states actually correspond to the discrete "wave patterns" which were possible with respect to the physical boundary conditions involved. Again the transitions between the states corresponded to discontinuous "quantum jumps" associated with a change in the total energy.

Another important consequence of the new theory was found by Heisenberg in 1927: it turned out that all electrons are in principle identical and indistinguishable, and that this fact leads to a new symmetry law and the so-called exchange phenomenon. By applying the new approach to the electron pair in the covalent bond of the hydrogen molecule, W. Heitler and F. London could in 1927 show that the energy of this bond is actually stabilized by the identity principle, and that the nature of Lewis's covalent bond hence depends essentially on a typical quantum-mechanical phenomenon, namely the exchange of electrons. This work opened the new field of "quantum chemistry." According to modern quantum theory, it should thus be possible to calculate the stationary states of a given molecule: the ground state, the various isomeric states, and all the excited states. The modern development has shown that the calculations involved may be of an immense order of magnitude, but the important thing here is the fundamental laws and principles.

C. Mutations as Quantum Jumps

In the 1930's, M. Delbrück (Delbrück *et al.*, 1935) pointed out the close analogy between the fundamental laws of genetics and of quantum theory: the immense stability of the hereditary substance in the chromosomes over thousands of years may indicate that it may be nothing but an immense molecule in a stationary state, and the mutations would then correspond to "quantum jumps" to isomeric forms.

Quantum-mechanical considerations of the type common in reaction kinetics imply that the probability for a quantum jump is roughly proportional to the quantity $\exp(-\Delta E/kT)$, where ΔE is the activation energy from the state under consideration up to the energy threshold leading over to other state, and kT is the average heat energy. Experimental data for mutations have shown at least a rough agreement with such a type of law, but we will return to this question later.

The Delbrück model has been discussed in greater detail by Schrödinger, in his Dublin lectures in 1943 and in his little book *What is Life?* (Schrödinger, 1945), and he points out that the gene in the chromosome fiber may suitably be called an *aperiodic crystal.* He emphasizes that physicists usually deal with periodic crystals and that, even if solid-state physics experimentally and theoretically is a highly complex and fascinating field where many important results have been achieved, the periodic structure is certainly plain and dull in comparison to an aperiodic crystal. "The difference in structure is of the same kind as that between an ordinary wallpaper in which the same pattern is repeated again and again in regular periodicity and a masterpiece of embroidery, say a Raphael tapestry, which shows no dull repetition, but an elaborate, coherent, meaningful design traced by the great master." For a study of the aperiodic crystal which is the material carrier of life, Schrödinger points to organic chemistry. Let us hence see what has happened on the chemical side.

D. Chemistry of the Nucleic Acids

In the beginning of the 19th century there had been a gap between organic chemistry and inorganic chemistry which was bridged by F. Wöhler's famous synthesis of urea in 1828 showing that "vital forces" from living cells were not absolutely necessary in producing organic compounds. At the same time as the cytologists started discovering the inner details of the cell, it became clear that the laws of ordinary chemistry were valid also within living matter.

The substances in the cytoplasm and in the cell nucleus were analyzed by the chemists and, in the latter, F. Miescher discovered in 1868 certain compounds which are now called "nucleic acids." Our knowledge of these compounds has been further enriched through the work by E. Hammarsten, T. Caspersson, and many others, and for a survey we would like to refer to the comprehensive book by Davidson (1960).

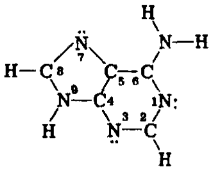
Through later research it has become clear that the nucleic acids are not confined to the cell nucleus but may be found also in the cytoplasm. Today, one distinguishes between two types of nucleic acids: deoxyribonucleic acids (DNA) and ribonucleic acids (RNA). The main constitutents of these compounds are aromatic nitrogen bases, pentose sugars, and phosphate groups. The nitrogen bases are of either purine or pyrimidine type. A schematic representation of these bases is given by Fig. 8,

Purine Pyrimidine

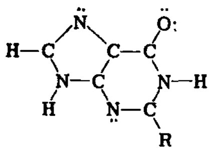
Fig. 8. The simplest nitrogen bases.

which also indicates the conventional numbering system of the ring atoms involved. In DNA, there are four main bases: adenine (A) and guanine (G) of purine type, and thymine (T) and cytosine (C) of pyrimidine type. RNA contains the same bases, but thymine is here replaced by uracil (U). There are also some small biological variations, and the detailed chemical structure of all the common bases is given in Fig. 9.

It should be observed that all the bases have the simple form of isomerism called "tautomerism" which may be characterized by moving a proton

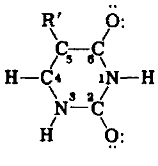


Adenine (A)



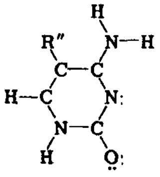
Guanine bases:

R —NH_2 Guanine (G)
R —H Hypoxanthine



Thymine bases:

R′ —CH_3 Thymine (T)
R′ —H Uracil (U)
R′ —Br Bromuracil (base analog)

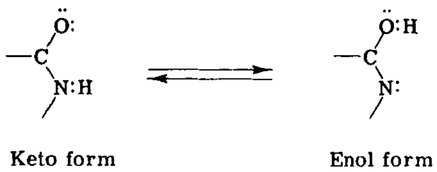


Cytosine bases:

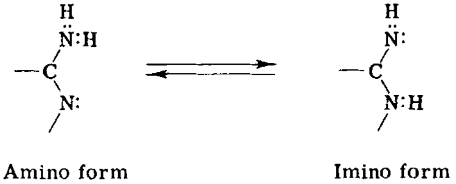
R″ —H Cytosine (C)
R″ —CH_3 Methylcytosine
R″ —Br Bromcytosine (base analog)

Fig. 9. Chemical structure of the common bases and some of the base analogs; in addition to the σ skeleton indicated by the figure, there is also a π electron cloud corresponding to the conventional double bonds. The double dots (:) indicate electron lone pairs which attract protons and will participate in hydrogen bonding.

from one electron pair to another. There are two main types, the keto-enol tautomerism:



and the amino-imine tautomerism:



In Fig. 9, the keto and amino forms have been considered as standard.

It was found rather early (see, e.g., Davidson, 1960) that the pentose sugar groups occurring in the nucleic acids were riboses of the following type shown in Fig. 10. In the deoxyribose, the oxygen attached to the 2 position in ribose has been removed.


(a) (b)

Fig. 10. The pentose sugar groups occurring in the nucleic acids: (a) β-ribose (RNA), (b) β-2-deoxyribose (DNA).

A nitrogen base may be condensed with a ribose or deoxyribose to form a *nucleoside*, and one distinguishes between ribonucleosides and deoxyribonucleosides. Adenosine, guanosine, cytidine, and uridine are the ribonucleosides formed from adenine, guanine, cytosine, and uracil, respectively, whereas thymidine is conventionally the deoxyribonucleoside formed from thymine. Experimental evidence has shown that in purine nucleosides the sugar group is attached in the N-9 position, whereas in the pyrimidine nucleosides the sugar is attached in the N-3 position. In both cases only the β configuration occurs in nature.

The phosphoric esters of the nucleosides are called *nucleotides*, and one distinguishes between ribonucleotides and deoxyribonucleotides. One has introduced the symbols MP, DP, and TP for the mono-, di-, and triphosphates, respectively, so that ATP means adenosine triphosphate. It should be observed, however, that the symbols are not unique, and that it is necessary to indicate the positions (2′, 3′, and 5′ in the ribonucleosides, for instance) where the phosphate groups are attached. A symbol without position indication is conventionally assumed to mean a compound with the phosphate group (or groups) attached to the 5′ position; see, for example, Fig. 11. The adenosine triphosphate (ATP) is famous in biochemistry as the main source of energy easily available in the cell; under release of energy the molecule is transformed into ADP and a phosphate group. It should finally be observed that a "d" in front of a molecular symbol indicates that one deals with a deoxyribonucleotide, as for instance dAMP, dGMP, ..., etc.

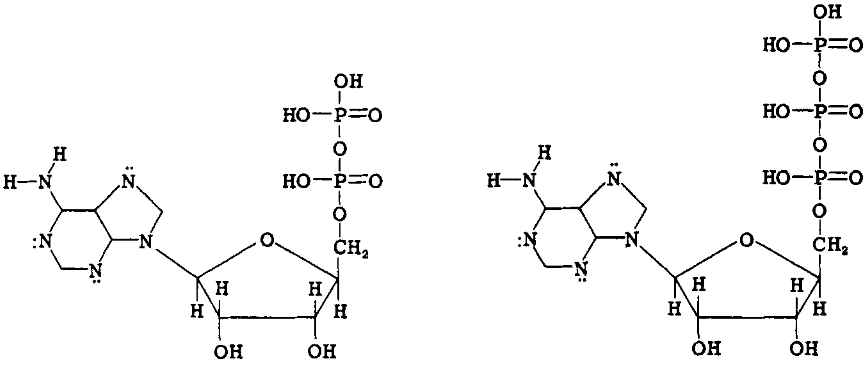
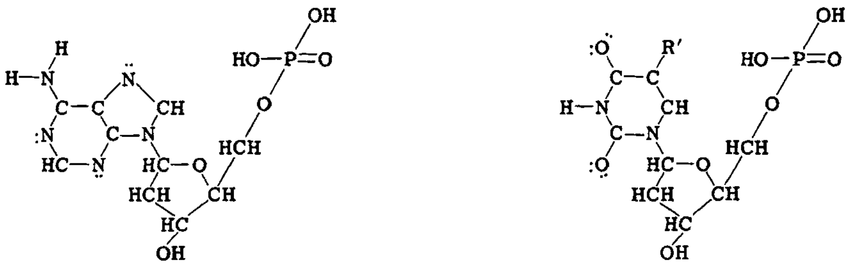
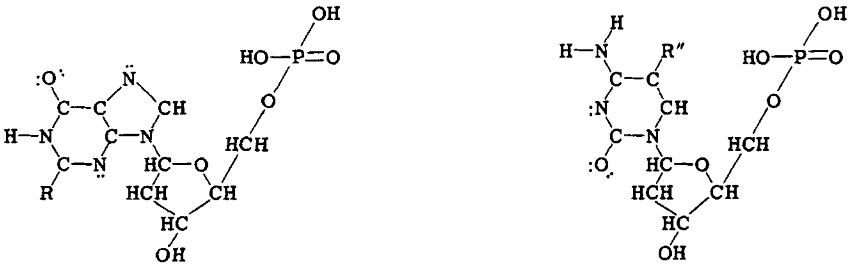

ADP ATP

Fig. 11. Two typical examples of nucleotides.

Experimental research has revealed (see Davidson, 1960) that the nucleic acids are *polynucleotides* formed from the monophosphates AMP, GMP, CMP, TMP, etc., obtained by linking the phosphate group of one nucleotide to the OH group of the sugar ring in the next nucleotide, hence leading to a chain structure. The chemical structure of the main mononucleotides are given in Fig. 12, and the chain structure is indicated in Fig. 13. As shorthand for such a compound, the notation given in Fig. 14



dAMP dTMP



dGMP dCMP

Fig. 12. The nucleotides which are the building stones of DNA; the explanation of the radicals R, R′, and R″ is given in Fig. 9. Note that all the sugar links are of β type.

has been used. Each horizontal line of Fig. 14 indicates a specific nucleotide (with the positions 1′, 2′, 3′, 4′, and 5′ in the pentose chain) and P the phosphate link between the successive nucleosides; the bases involved are indicated by the letters to the left. Until the end of the 1940's, one used to believe that the nucleic acids contained equimolecular amounts of all of the four main bases and hence showed a tetranucleotide structure of a high degree of periodicity. This hypothesis is now completely abandoned after a most remarkable development in the field of genetics.

E. DNA as Carrier of Genetic Information

One of the most interesting discoveries within the field of immunology was the *transformation principle* found by F. Griffith in 1928 in his work concerning the properties of pneumococci. He could show that non-virulent living pneumococci could obtain virulent properties by adding extract from the dead cells of virulent species, which indicated that the

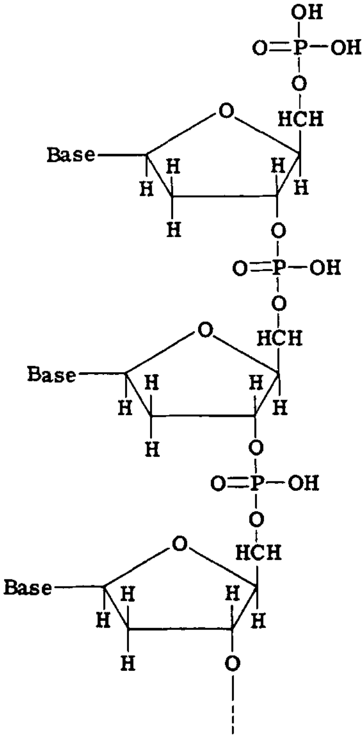


Fig. 13. Chain structure of a polydeoxyribonucleotide. The phosphate linkage goes between the 5′ position in one sugar to the 3′ position in the next sugar.

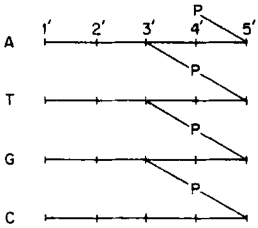


Fig. 14. Short-hand notation to indicate the chain structure of a polynucleotide.

transformation was achieved by some chemical compound or compounds characteristic for the virulent type. After many years of research, Avery and his co-workers could in 1944 show conclusively that the compound involved in the transformation was DNA. They could also show that transformations may quite generally be induced in a culture of organisms by adding DNA extracted from a mutant, and that DNA hence seems to be the essential carrier of the genetic information. This result seemed like a shock to most scientists involved, since one could hardly believe that a strictly periodic tetranucleotide structure could carry any information of the type required.

The solution of this mystery came in 1950 through the work of Chargaff and his collaborators who studied the molar proportions of the nucleotide bases in DNA and RNA by means of paper chromatography and could show that the tetranucleotide hypothesis was completely false (Chargaff, 1950, 1955; Chargaff *et al.*, 1950). Instead they found that, in DNA, the two bases adenine and thymine occur with the same molar content and that the same is true for the two bases cytosine and guanine, so that

$$\frac{\mathrm{A}}{\mathrm{T}} = \frac{\mathrm{C}}{\mathrm{G}} = 1. \tag{II.1}$$

On the other hand, the ratio of adenine to guanine varies greatly in DNA from different sources. For RNA, they found the somewhat looser rule that the molar content of adenine and cytosine together equals the molar content of guanine and uracil, so that

$$\mathrm{A} + \mathrm{C} = \mathrm{G} + \mathrm{U}. \tag{II.2}$$

Through this work, it became clear that the DNA molecule may have the character of the "aperiodic solid" discussed by Schrödinger in 1943, and that the genetic message may be contained in the four-letter code-script involving the four bases A, T, G, and C along a polynucleotide chain. The result agrees very well with Schrödinger's idea that the genetic code should consist of a well-ordered association of atoms having the stability of a molecular arrangement and with his statement: "Indeed, the number of atoms in such a structure need not be very large to produce an almost unlimited number of possible arrangements."

III. Stereostructure of DNA and the Connection with Its Biological Functioning

A. The Wilkins–Watson–Crick Model of DNA

The stereostructure of the nucleic acids had for a long time been an important problem, and its study was further intensified after the discovery that the DNA molecule was the carrier of the genetic information in the cell. The rough structure of the polynucleotide chain had been revealed by purely chemical means, see Fig. 3, and refined physicochemical methods had shown that the molecule was about 20 Å thick and many thousands of angstroms in length and hence had a remarkable fiber structure.

The details of the atomic structure could now be investigated by means of X-ray analysis. This technique originally developed for crystals by von Laue and the Braggs had been used successfully by Pauling and his collaborators in 1951 to study the proteins and had led to the discovery of the helical structures and the importance of the hydrogen bonding (Pauling and Corey, 1951; Pauling *et al.*, 1951). By means of X-ray analysis of the polynucleotides, Furberg (1950, 1952) had determined the dimensions of the sugar-phosphate backbone and the bases, and he suggested that the structure would be a single helix with a radius of approximately 10 Å. Further X-ray data for DNA were obtained by Wilkins and his collaborators and, on the basis of these results, two groups independently suggested that DNA would be a *double helix* having the sugar-phosphate chain in the outside of the helix, namely, Wilkins *et al.* (1953a, b) and Watson and Crick (Watson and Crick, 1953; Crick and Watson, 1954). The latter also made a careful study of the hydrogen bonding in the structure and its biological implications, and this has been of fundamental importance for the entire development in this field during the last decade.

According to this stereomodel, the DNA molecule consists of a double helix formed by two sugar-phosphate chains of the classical type (see Fig. 13) held together by a sequence of base pairs joined by (at least) two hydrogen bonds (see Fig. 15). Each base pair is approximately a planar conjugated system perpendicular to the long axis of the helix, and it consists of a purine base and a pyrimidine base attached to the sugar groups in the conventional way. The bases are assumed to exist in their keto and amino forms, respectively, and the hydrogen bonding is then unique so that adenine is always attached to thymine (A—T) and cytosine to guanine (C—G) in complete agreement with Chargaff's experimental result, expressed in Eq. (1); see also Fig. 16.

The double helix has a diameter of approximately 20 Å; each helix makes a full turn after an axis translation of 34 Å containing ten base pairs, and there is hence a translation of 3.4 Å per base pair and a rotation of 36°. Since the sugar linkage is solely of the β type, both helices are necessarily right-handed and the sugar-phosphate chains run in opposite directions.

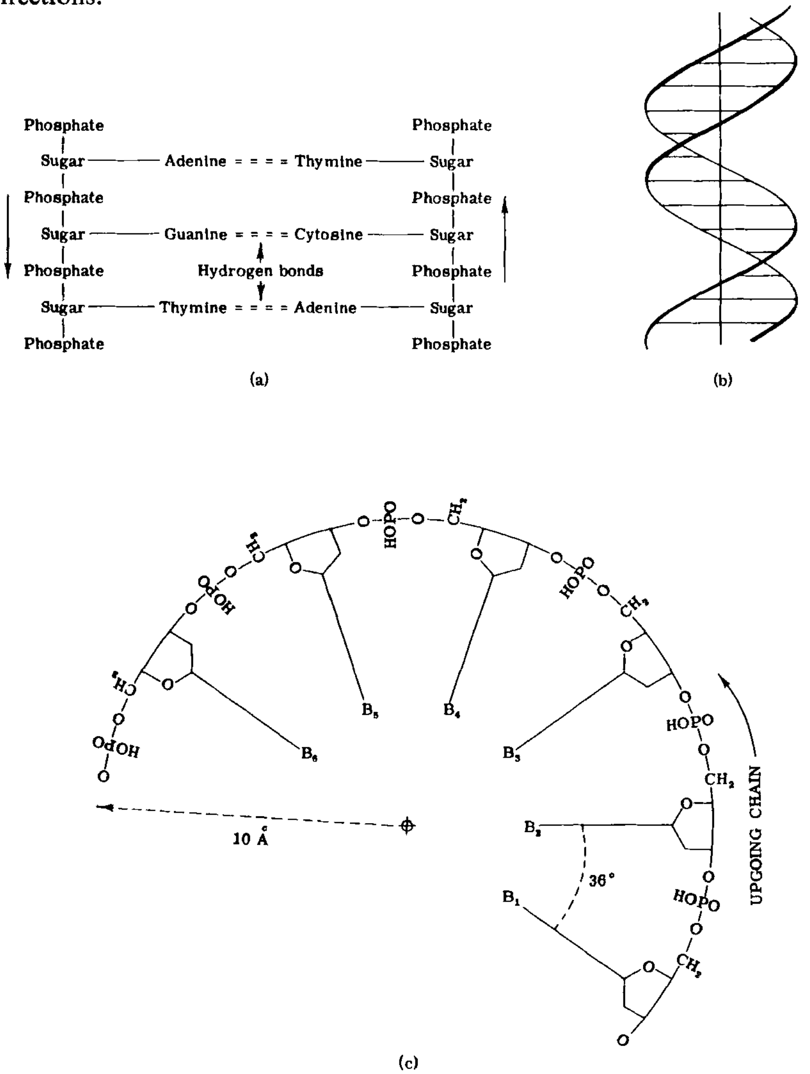
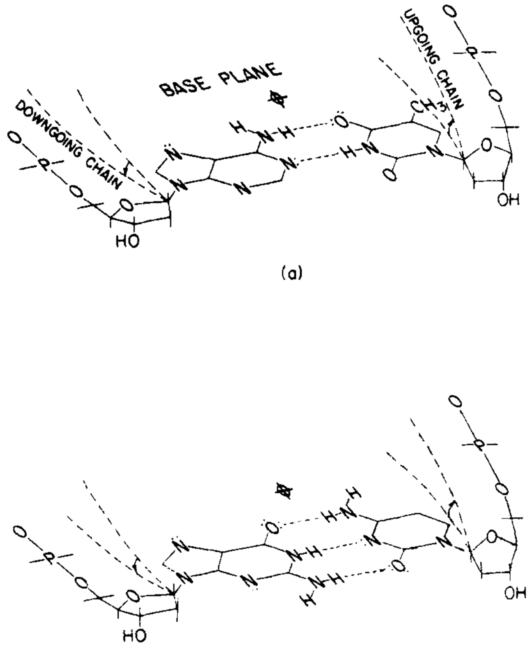


Fig. 15. Schematic structure of DNA: (a) Linear representation of the DNA molecule. (b) Double helix of DNA. (c) One of the sugar-phosphate chains seen from the top of the helix axis, diameter approximately 20Å.



(a)

(b)

Fig. 16. The normal base pairs occurring in DNA: (a) A—T base pair, (b) G—C base pair.

B. Complementarity and DNA Duplication

The Watson–Crick model implies that each base B has its specific *complementary* base B′ and, if one strand has the base sequence B_1, B_2, B_3, B_4, ..., the other strand has the complementary sequence B_1', B_2', B_3', B_4', ... It should be observed that the complementarity concept depends not only on the hydrogen bonding but also on steric factors: a short pyrimidine base can join only a long purine base, and vice versa, to fit the dimensions of the double helix.

The genetic message in DNA is contained in each one of the two strands in complementary form. Before the cell division, the DNA molecules should in some way be duplicated and, according to Watson and Crick, the double helix starts unwinding at the same time as each strand starts building its own complement giving rise to two identical DNA molecules containing the original genetic information (see Fig. 17). The basic principle is exceedingly simple, but the actual mechanism involved may be complicated; this problem will be further discussed in a following section.

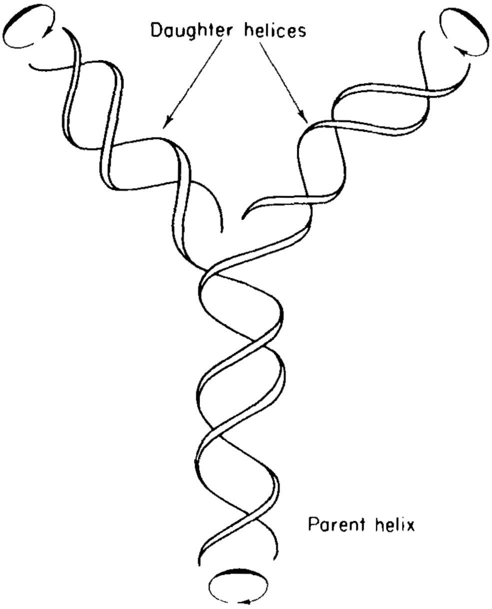


Fig. 17. Replication of DNA according to Watson and Crick; winding mechanism after M. Delbrück and G. S. Stent.

C. Complementarity, Tautomerism, and Mutations

It is clear that, according to Watson–Crick's model, the complementarity property of the bases is essential for both the stability of the genetic code and the duplication mechanism.

In order to study the concept of complementarity in greater detail, it is worthwhile to consider the parts of the bases which take part in the formation of the hydrogen bonds. These bases are illustrated in Fig. 18 (see also Fig. 16). Writing the bases in this particular way, we can introduce a shorthand notation for the "proton-electron pair" code see Fig. 19 (a)]. The bases A and C have *equivalent codes* (with respect to the upper two positions) and the same is true for T and G. In these figures, we have emphasized the electron lone pairs and the protons H, and we note that a hydrogen bond is essentially a proton shared between two electron lone pairs associated with different atoms.

In addition to the normal forms, we will now also consider the previously mentioned tautomeric forms obtained by moving a proton from the upper lone pair to the middle one, or vice versa. Denoting the imine forms of A and C by A* and C*, respectively, and the enol forms of T and G by

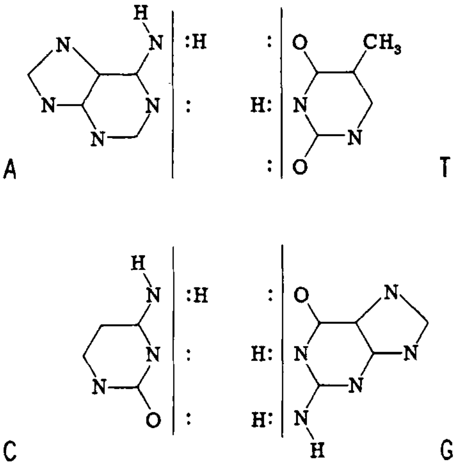


Fig. 18. The nucleotide bases occurring in DNA and their templates for hydrogen bonding; note that H represents a proton and the symbol (:) an electron pair according to Lewis.

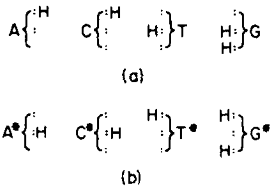


Fig. 19. (a) The proton-electron pair code of the normal nucleotide bases. (b) The proton-electron pair code of the rare tautomeric forms of the nucleotide bases.

T* and G*, respectively, we obtain the "proton-electron pair" codes shown in Fig. 19 (b). From a study of the hydrogen bonds, it is now clear that A* will no longer combine with T but with C, etc., so that one obtains the combinations

$$A^* - C, \quad A - C^*, \quad G^* - T, \quad G - T^*.$$

This means that the complementarity between the bases is completely changed, and the movement of a single proton within a base will in this way influence the genetic message and introduce an error at the first cell duplication, according to the following scheme:

original sequence:	A G T C A T T G C A
tautomeric change:	A G T* C A T T G C A
complementary sequence:	T C G G T A A C G T
new sequence:	A G C C A T T G C A

The general diagram below gives a comparison between the normal and the tautomeric replication of the single bases with the complementary base in the middle.

Normal	Tautomeric
A—T—A	A*—C—G
T—A—T	T*—G—C
G—C—G	G*—T—A
C—G—C	C*—A—T

It is illustrative to compare the replication of a normal base pair with the replication of a pair containing a tautomeric form. Figure 20 roughly shows that an A—T* pair in part of the "offsprings" gives rise to a G—C pair. Such a genetic error undergoes then "biological amplification" by the factors 2, 4, 8, 16, ..., and may soon become recognizable.

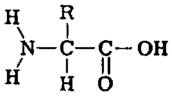
Fig. 20. Normal and tautomeric replication of a base pair. The latter process is in reality more complicated (see Figs. 73 and 74).

Watson and Crick (1953) have suggested such a mechanism to explain the *mutations*: "Spontaneous mutations may be due to a base occasionally occurring in one of its less likely tautomeric forms." We note that this is in complete accordance with the general idea expressed by Delbrück and discussed by Schrödinger. The "quantum jumps" corresponding to the mutations would then be associated with *proton transfer* within the DNA molecule, and the quantum-mechanical mechanism involved in such a process will be discussed in a later section.

D. Structure of the Proteins

The problem is now how the genetic information contained in the DNA molecule determines the biological properties of the species and individual under consideration. Each species is characterized by its proteins, and of particular importance are the enzymes which catalyze the entire metabolism. The proteins are essentially linear structures built up from twenty amino acids, and the biochemical properties are determined by the sequence of the amino acids. This sequence must, of course, ultimately be determined by the base sequence in DNA, and the question of the connection between these two linear arrangements gives rise to the *coding problem.*

Let us start with a brief review of the chemical background. An amino acid is characterized by the structure:



$$\begin{array}{c} H \\ H \end{array}\!\!>N-\overset{R}{\underset{H}{C}}-\underset{\underset{O}{\|}}{C}-OH$$

where R indicates one of the possible "residues." The twenty residues occurring in nature are listed in Fig. 21. Two amino acids may be joined

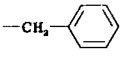
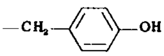
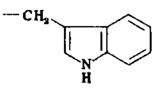
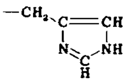
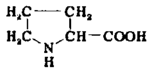
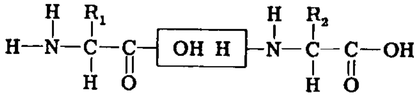
	R			R	
Glycine	$-H$	Gly	Phenylalanine		Phe
Alanine	$-CH_3$	Ala			
Valine	$-CH(CH_3)_2$	Val			
Leucine	$-CH_2CH(CH_3)_2$	Leu	Tyrosine		Tyr
Isoleucine	$-CH(CH_3)CH_2CH_3$	Ileu			
Serine	$-CH_2OH$	Ser			
Threonine	$-CHOH-CH_3$	Thr	Tryptophan		Try
Cysteine	$-CH_2SH$	CySH			
Methionine	$-CH_2CH_2SCH_3$	Met			
Aspartic acid	$-CH_2COOH$	Asp	Histidine		His
Asparagine	$-CH_2CONH_2$	Asp NH_2			
Glutamic acid	$-CH_2CH_2COOH$	Glu			
Glutamine	$-CH_2CH_2CONH_2$	Glu NH_2			
Lysine	$-(CH_2)_3CH_2NH_2$	Lys	Proline		Pro
Arginine	$-(CH_2)_3-NH-C(NH_2)=NH$	Arg			

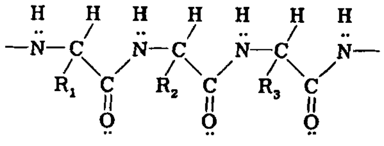
Fig. 21. The twenty amino acids found in proteins.

by the *peptide bond* suggested independently in 1902 by Fischer and Hofmeister:



$$H-\overset{H}{N}-\overset{R_1}{\underset{H}{C}}-\underset{\underset{O}{\|}}{C}-\boxed{OH\ H}-\overset{H}{N}-\overset{R_2}{\underset{H}{C}}-\underset{\underset{O}{\|}}{C}-OH$$

and a "polypeptide" has the structure:



A great deal of research in biochemistry has revealed that the proteins are essentially giant polypeptides and, by ingenious methods, one has determined the sequence of the residues R_1, R_2, R_3, ... in some of the most important enzymes. The X-ray analysis by Pauling *et al.* (1951) has shown that some of the proteins, like α-keratin, have a helical structure and are further stabilized by hydrogen bonds between neighboring chains (see Fig. 22). All these helices seem to be right-handed in an absolute sense, and the residues are pointing out from the cylinder. Other proteins, like myoglobin, have a much more complicated stereostructure (Crick and Kendrew, 1957; Kendrew, 1959; Berg, 1961).

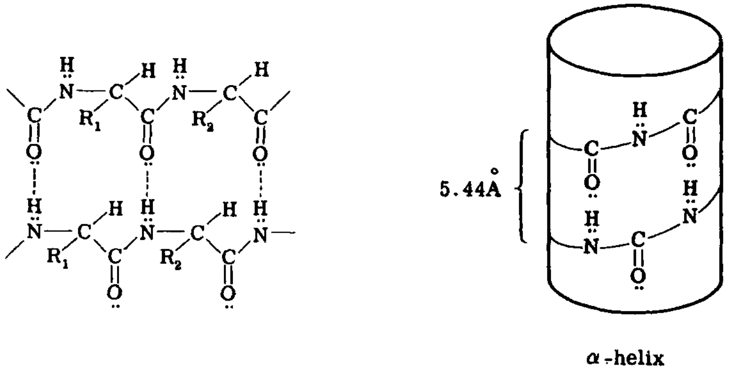


Fig. 22. Stabilization of a polypeptide α-helix through hydrogen bonding.

E. Role of DNA and RNA in Protein Synthesis

Let us now return to the coding problem. Through the work by Caspersson (1942, 1950) and others (see, e.g., Brachet, 1950) it had become clear that cells engaged in protein synthesis had a large content of RNA also in the cytoplasm. It is now clear that RNA serves as an intermediate between DNA and the proteins, so that DNA regulates RNA which in turn controls the protein synthesis. Since even the enzymes which catalyze the DNA duplication are produced in this way, one obtains the results illustrated in Fig. 23, which may be characterized as the "growth cycle." Actually, there is a "feedback" mechanism at several of the other links in the diagram.

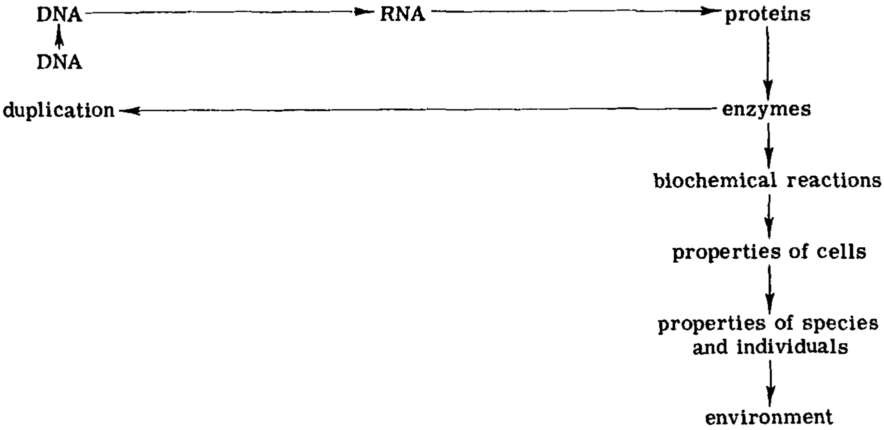


Fig. 23. Schematic diagram of the control mechanism of the cell.

Studies of the cytoplasm have revealed that protein synthesis takes place in or on small bodies called *ribosomes*, and that it is regulated by RNA. However, it has turned out that RNA is not a single type of molecule but a complex of molecules with various biochemical functionings (see, e.g., Crick, 1958; Hoagland, 1959; Brachet, 1960). The actual genetic information seems to be contained in an enormously long, linear, single-stranded molecule called *messenger* RNA which has originally in some way picked up the information at the DNA in the cell nucleus. The dimensions are such that the ribosome is a comparatively small particle in the form of a "ball" or ring gliding along the giant RNA chain.

For the linear arrangement of the amino acids in the protein, Crick (1958) and Hoagland (1959) have introduced the so-called *adapter hypothesis* according to which the amino acids are picked up by small pieces of RNA molecules called "soluble RNA" or "sRNA." "The position of a particular amino acid is then determined not by the amino acid itself but by the hydrogen bonding between the messenger RNA template and a complementary nucleotide sequence in the sRNA carrying the amino acid." Experimental evidence seems to verify this idea and the role of both messenger RNA (Fraenkel-Conrat *et al.*, 1962) and sRNA (Chapeville *et al.*, 1962).

F. Coding Problem

The problem how the four-letter message in the base sequence B_1, B_2, B_3, ... in DNA is transferred to the twenty-letter message R_1, R_2, R_3, ... in the proteins is one of the key problems in protein synthesis. Here we will leave the biochemical aspects somewhat aside and instead consider the mathematical problems connected with the transfer of a certain amount of information. It is clear that, since $4 < 20$, a single letter in the base sequence cannot determine a letter in the peptide sequences and, since $4^2 < 20$, the same is true also for pairs of letters. It seems hence necessary that the base sequence in some way is "read" in at least *triplets* of letters but, since $4^3 = 64$, one has then to deal with a code which is either degenerate, partly senseless, or both.

The coding problem was formulated by Dounce (1952), but the first real attempt to solve it was made by Gamow (1954a, b). He had observed that, in the double helix of DNA, there were "holes" in the shallow groove with about the same interspacing as the amino acids in the proteins. Each "hole" was surrounded by four base pairs, but, since the middle pair consisted of complementary bases, there were actually only three bases involved [see Fig. 24(a)]. The code was hence a "triplet code" and, by assuming that it would further be symmetric with respect to the reading direction so that $B_1B_2B_3$ would be equivalent to $B_3B_2B_1$, Gamow could show that it led to twenty independent combinations. These are indicated in Fig. 25(a), where, for simplicity, we have denoted the bases by the

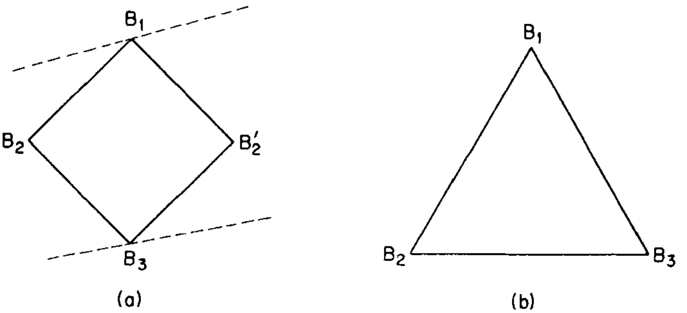


Fig. 24. (a) Gamow "hole" in diamond code. (b) Gamow "hole" in triangular code.

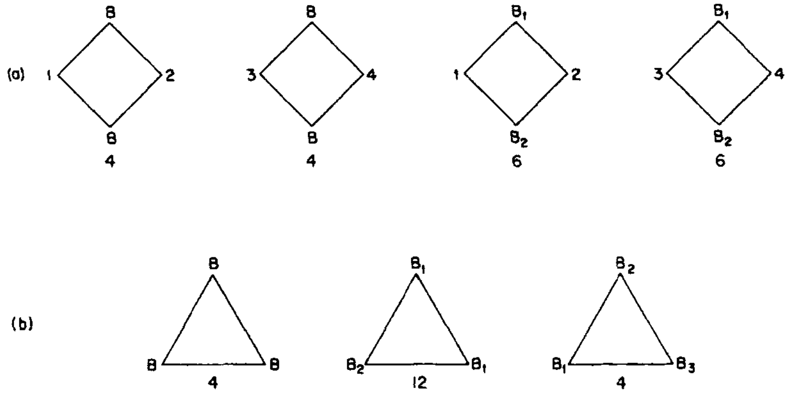
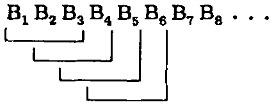


Fig. 25. (a) Number of possibilities in diamond code. (b) Number of possibilities in triangular code.

numbers 1, 2, 3, 4 and assumed that 1–2 and 3–4 are complementary pairs. This code has been called the "diamond code" depending on the form of the figures, and since that base sequence is "read" in the following way:

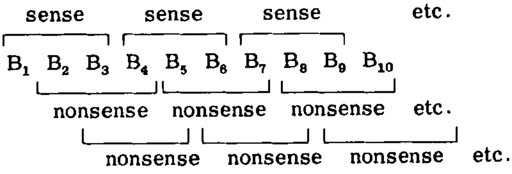


it is characterized as an "overlapping triplet code." Gamow realized that these overlappings, i.e., the fact that two neighboring holes have two common nucleotides, would have important consequences and lead to a partial correlation between neighboring amino acids in the proteins. If $B_1B_2B_3$ codes the amino acid α and $B_2B_3B_4$ the amino acid β, the sequence $B_1B_2B_3B_4$ would code the dipeptide $\alpha\beta$, However, since there are $4^4 = 256$ possible quadruples $B_1B_2B_3B_4$ but $20^2 = 400$ dipeptides, it is clear that there would be many dipeptides which could never occur.

Another overlapping code suggested by Gamow and his collaborators (Gamow and Ycas, 1955; Gamow *et al.*, 1956) was the "triangular code" based on the idea that the triplet $B_1B_2B_3$ may be equivalent with all other triplets obtained by rotation and/or reflexion [see Fig. 24(b)]. Even this code leads to exactly twenty independent triplets, as indicated in Fig. 25(b).

During the years following Gamow's hypothesis, there was an intense research as to overlapping codes and the question of correlation in dipeptides. However, no experimental evidence for such a correlation could be found, and finally Brenner (1957) showed that there seemed to be a sufficiently large number of dipeptides in nature to prove the impossibility of all overlapping triplet codes.

In 1957, Crick and his collaborators introduced another interesting development through the "comma-less code" meaning that there would only be one division of the base message into triplets which had sense, whereas all other divisions would be senseless:



On the basis of this simple rule, they could show that there existed such "comma-less" codes. If the triplet BBB is assumed to code the amino acid α, the sequence BBB BBB would code the dipeptide $\alpha\alpha$ but would also contain two more triplets which ought to be nonsense; hence the entire triplet should be rejected. This leaves $64 - 4 = 60$ triplets at our disposal. Let us then consider a sense triplet $B_1B_2B_3$ which is assumed to code the amino acid β. The sequence $B_1B_2B_3B_1B_2B_3$ codes the dipeptide $\beta\beta$ and contains the senseless triplets $B_2B_3B_1$ and $B_3B_1B_2$, i.e., the triplets obtained from the original one through cyclic permutations. Since $60:3 = 20$, it is clear that the maximum number of "sense triplets" cannot exceed the magic number twenty. Crick *et al.* show by an example that there exist actually exactly twenty sense triplets:

Type I:	$12\begin{smallmatrix}1\\2\end{smallmatrix}$		$\begin{smallmatrix}1\\2\end{smallmatrix}3\begin{smallmatrix}1\\2\\3\end{smallmatrix}$		$\begin{smallmatrix}1\\2\\3\end{smallmatrix}4\begin{smallmatrix}1\\2\\3\\4\end{smallmatrix}$	
number of possibilities:	2	+	6	+	12	=20

In each group with a fixed middle figure, one may take all combinations of the first and the last figures to form sense triplets. The "comma-less" code has a great physical advantage, since it implies that only "sense triplets" can catch an amino acid which comes then from the very beginning into the correct place.

All possible types of comma-less codes have now been systematized

(Freudenthal, 1958; Golomb *et al.*, 1958) and four more types should be mentioned:

Type II:	211	122		${}^{1}_{2}3\overset{1}{\underset{3}{2}}$	${}^{1}_{3}2\,4^{\overset{1}{2}}_{\underset{4}{3}}$
Type III:	${}^{1}_{3}2^{1}_{2}$		${}^{1}_{2}3^{1}_{3}$		${}^{1}_{3}2\,4^{\overset{1}{2}}_{\underset{4}{3}}$
Type IV:	12^{1}_{2}		$\begin{matrix}1&&1\\2&3&2\\4&&3\end{matrix}$		$\begin{matrix}1&&1\\2&4&2\\3&&4\end{matrix}$
Type V:	121	211		$\begin{matrix}1&&1\\2&3&2\\4&&3\end{matrix}$	$\begin{matrix}1&&1\\2&4&2\\3&&4\end{matrix}$

With these types, the possibilities for constructing comma-less codes are extinguished.

The coding problem is still not fully solved, and there remains the important question whether there actually exists a "universal code," or whether the code could be different for various species. Using experimental evidence from amino acid sequences and properties of mutants, several authors have contributed to the solution of the problem (Levinthal, 1959; Ycas, 1960; Woese, 1961a, b; Zubay and Quastler, 1962; Roberts, 1962; Benzer and Champe, 1962). By now even the comma-less codes seem to be abandoned, and the situation has been summarized by Crick *et al.* (1961) in the following way: the code seems to be a degenerate non-overlapping triplet code which is read from a fixed starting point without special "commas." The word "degenerate" means that here there are several triplets which may code one and the same amino acid.

Recently the entire coding problem has been moved closer to a solution through a remarkable development in biochemistry. Extending the techniques developed by Kornberg and by Ochoa for the synthesis of polynucleotides, Nirenberg and Matthei (1961) succeeded in using a simple polynucleotide of the type UUUUU ... as a template in the biosynthesis of polypeptides. Since the result was a polyphenylalanine peptide, it seems hence as if the group UUU would code phenylalanine. Further studies of a polynucleotide containing the bases U and G in the ratio 5:1 showed that the corresponding polypeptide, in addition to phenylalanine, contains the amino acids leucine, valine, glycine, and tryptophan. The proportions are such that it seems likely that the coding letters for leucine and valine are UUG and for glycine and tryptophan UGG. This development has led to a new approach in the coding problem, which will be treated in detail in a later section.

G. Viruses

The study of viruses started with the discovery that certain diseases could be produced by particles which would pass through filters stopping bacteria and which hence would have much smaller dimensions. After several decades of intense research, we know today that a virus consists essentially of a DNA molecule or RNA molecule in a protein "overcoat." When a virus approaches a cell, it will throw off its protein overcoat, and the DNA or RNA molecule will enter through the cell membrane. The virus DNA or RNA molecule will then take over the control of the metabolism of the "host" cell, and it will start replicating itself until most of the cell material is exhausted. The new DNA or RNA molecules will then be wrapped into protein overcoats so that, when the cell membrane bursts, they are ready to enter the environment and approach new host cells.

One distinguishes between four types of viruses: bacterial, animal, plant, and insect viruses, of which the first type (the "bacteriophages") has been of particular importance in the study of DNA and RNA as hereditary substances. One has found that the single-stranded virus RNA molecule has the same power of replication as the double-stranded DNA molecule, and that, in exceptional cases, there seems to exist also a single-stranded DNA molecule. However, in all cases, one believes that the replication uses a template mechanism built on Watson and Crick's complementarity idea. For a survey of the rich literature on viruses, we would like to refer to a few selected papers (Hershey and Chase, 1952; Pollard, 1953; Fraenkel-Conrat *et al.*, 1957, 1962, Fraenkel-Conrat, 1959; Burnet and Stanley, 1959).

Whether the viruses are to be considered as "living" or "dead" matter is ultimately a question of semantics. A virus may be considered as a giant molecule or a molecular complex in a stationary state which may remain the same for a very long time and has all the characteristics of dead organic matter. However, when the virus is brought into interaction with a host cell, it starts replicating itself and shows thus the most fundamental property of "living" matter.

IV. Properties of the Hydrogen Bond

Since the Watson–Crick model essentially utilizes the hydrogen bond in the definition of the complementarity between the nucleotide bases, it may be worthwhile to study the properties of this bond in greater detail.

Chemical experience has shown that a hydrogen atom attached to an electronegative atom in a molecule may also be attracted to another electronegative atom in a different molecule, in this way leading to a "hydrogen bond" between the two molecules. Sometimes there is also an internal hydrogen bond between two electronegative atoms within the same molecule. The atoms which form the strongest hydrogen bonds are, in order after decreasing strength: fluorine, oxygen, and nitrogen, whereas weak bonds are formed by chlorine and carbon. Experimentally the properties of the hydrogen bonds have been studied extensively and, for a survey, we would like to refer to Pimentel and McClellan (1960) and to the proceedings from the 1957 conference (Hadzi, 1959).

A. Electron-Proton Formulation of the Hydrogen Bonding

In order to investigate the properties of the hydrogen bond, one has to understand the electronic structure of the atoms involved. According to the idea of hybridization introduced by L. Pauling and J. C. Slater in 1932, the orbitals of the valence electrons are superimposed or hybridized to form orbitals having a particularly high electron density in specific directions corresponding to directed valency. By superposition of an s orbital and a p_k orbital, one obtains a hybrid (shown in Fig. 26):

$$h = \alpha s + \beta p_k \qquad (\alpha > 0, \quad \beta > 0),$$

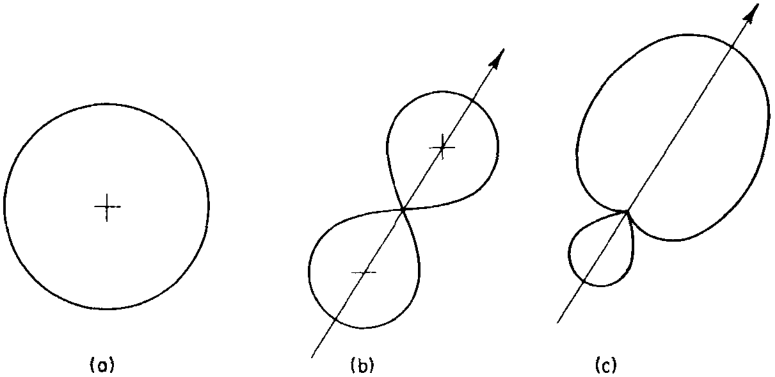

(a) (b) (c)

Fig. 26. Schematic diagrams of angular distributions of: (a) s function, (b) p function, and (c) hybrid.

which has its maximum density in the k direction. The normalization condition $\langle s|s\rangle = \langle p_k|p_k\rangle = \langle h|h\rangle = 1$ gives $\alpha^2 + \beta^2 = 1$. The ratio $n = \beta^2/\alpha^2$ indicates the amount of p character relative to the amount of s character, and the hybrid is conventionally denoted by the symbol sp^n, where n does not necessarily have to be an integer.

If there are two hybrids, h_1 and h_2, which are orthogonal in Hilbert space so that $\langle h_1|h_2\rangle = 0$, one has

$$\langle h_1|h_2\rangle = \alpha_1\alpha_2 + \beta_1\beta_2\langle p_1|p_2\rangle = 0.$$

Since $\langle p_1|p_2\rangle = \cos\theta_{12}$ gives the cosine for the geometrical angle θ_{12} between the two hybrids, this leads to

$$\cos\theta_{12} = -\frac{1}{\sqrt{n_1 n_2}} \tag{IV.1}$$

which relation gives the connection between the geometrical structure and the amount of hybridization.

If there are four mutually orthogonal hybrids sp^{n_1}, sp^{n_2}, sp^{n_3}, and sp^{n_4}, the total amount of s and p character has to be used up which leads to the auxiliary condition:

$$\frac{1}{n_1+1} + \frac{1}{n_2+1} + \frac{1}{n_3+1} + \frac{1}{n_4+1} = 1. \tag{IV.2}$$

The necessary and sufficient condition that three hybrids should be in the same plane is simply

$$\frac{1}{n_1+1} + \frac{1}{n_2+1} + \frac{1}{n_3+1} = 1,$$

and it is then easily seen that the fourth hybrid must be a pure p orbital perpendicular to the plane (see Fig. 27). The coplanar hybrids are said to

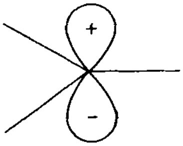


Fig. 27. Schematic diagram of three hybrids in a plane plus a $2pz$ function.

form the σ part of electronic structure, whereas the p orbital forms the π part. The σ and π orbitals are characterized by being symmetric and antisymmetric, respectively, with respect to reflections in the plane.

A typical example of a symmetric planar structure is the benzene molecule C_6H_6, where the angles are 120° and the carbon atoms have sp^2 hybridization (see Fig. 28). Three valence electrons from each carbon atom participate in the single bonds in the σ skeleton, whereas the fourth valence electron enters one of the molecular orbitals formed by linear superposition of the $2p_z$ orbitals of the individual carbon atoms and contributes to the π electron cloud. Let us now assume that we could carry out a thought experiment consisting of dropping one of the protons around the ring into one of the carbon nuclei giving rise to a nitrogen nucleus. This would lead to the formation of a pyridine molecule (see Fig. 29) characterized by having an *electron lone pair* sitting in a sp^2 hybrid on the nitrogen atom. The number of π electrons will remain unchanged

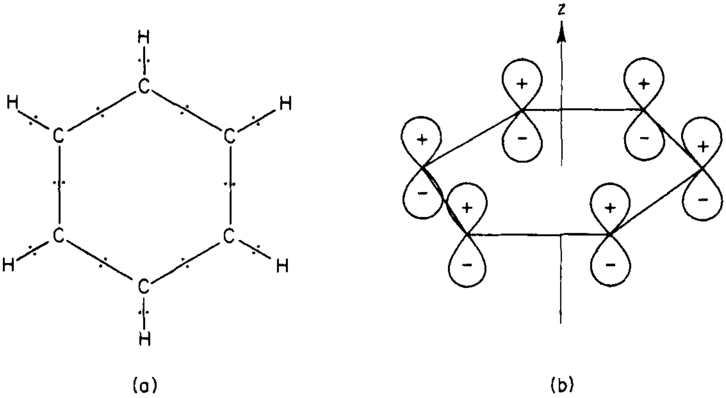


Fig. 28. The benzene molecule C_6H_6: (a) σ-skeleton, (b) π electron cloud.

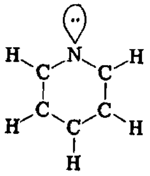
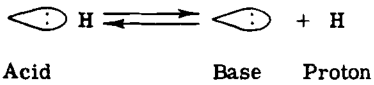


Fig. 29. Pyridine molecule.

and equal to 6. It is clear that the extra positive charge on the N atom will attract the mobile π electrons and serve as an "electron trap." Even the lone pair will be influenced, but the main structure will still be two electrons in a σ orbital pointing far out in space from the ring. This lone pair will attract every proton (or positive group) in the neighborhood, and this tendency of pyridine to try to catch a proton and become a pyridinium ion $C_5NH_6{}^+$ gives it the characteristic "base" character.

In general, protons are usually attached to molecules by means of electron pairs and, in many cases, the bonding is so strong that it is very hard to remove the protons. However, there is a large class of highly important chemical compounds in which the easy removal or attachment of protons is one of the fundamental chemical properties, namely the acids and the bases, which are essentially characterized by the reaction:



The main feature of the base is hence that it has at least one electron lone pair available, whereas the acid is characterized by having "loose" protons which are more or less easily removed. Of course, both properties may occur in one and the same molecule.

If there are several molecules having such electron lone pairs in a system, there may be a competition to catch the protons in the environment which leads to the formation of the above-mentioned hydrogen bonds. In this type of formulation, a hydrogen bond is characterized as a *proton shared between two electron lone pairs* (see Fig. 30).

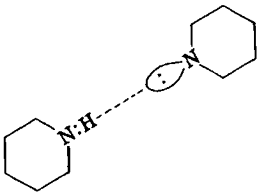
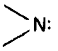


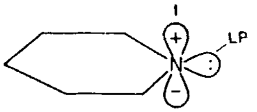
Fig. 30. Hydrogen bond between two pyridine molecules.

The following figures indicate the approximate electronic structure of nitrogen and oxygen atoms in some groups commonly occurring in conjugated systems.

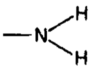
Nitrogen in:

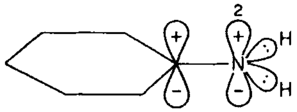
(1) ring:



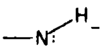


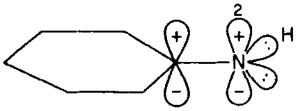
(2) amino group:





(3) imino group:

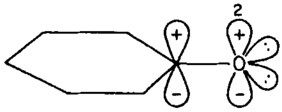




Oxygen in:

(1) keto group:

$-C=O^-$



(2) enol group:

$-C-OH$

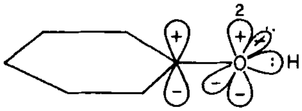


Table I gives a survey of the electron distribution of C, N, O, and F commonly occurring in these groups.

TABLE I

Atom	Valence electrons	Group	σ electrons		π electrons
			Single bonds	In lone pairs	
C	4	Ring	3	—	1
N	5	Ring	2	2	1
	5	Amino	3	—	2
N^-	6	Imino-	2	2	2
O	6	Enol	2	2	2
O^-	7	Keto-	1	2 + 2	2
F	7		1	2 + 2	2

The number in the last column gives the number of π electrons contributed to the mobile electrons. However, since part of these electrons in attached groups migrate into the ring, the actual π charge is always less than 2. This table gives a rough idea of why the strength of the hydrogen bond goes $F > O > N \gg C$, but there are many exceptions to this simple rule.

A substance which is highly characterized by hydrogen bonding is water, where the oxygen atom has at least one electron lone pair (see Fig. 31) easily available, and the water molecule then has the tendency to try to catch a proton and form the hydronium ion H_3O^+. This proton may be taken from another water molecule, which then forms the hydroxyl ion OH^- with *two* electron lone pairs easily available. The competition to catch protons between the water molecules leads to the effect of hydrogen bonding which is highly essential for the properties of the fluid (Lennard-Jones and Pople, 1951; Pople, 1951).

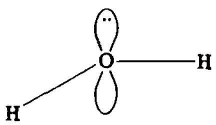
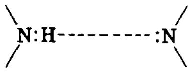
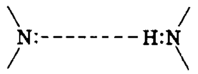


Fig. 31. Schematic structure of the water molecule.

In each hydrogen bond, the proton has two equilibrium positions—one close to each one of the two electron lone pairs involved:





If these equilibrium positions are equivalent, one can expect that, under certain conditions, the proton may jump from one position to another. Since such a jump will influence the gross electric neutrality of the entire environment, it may induce other proton jumps so that the final effect will be a collective phenomenon. This process in the ice crystal has been studied by Pauling (1939) and he shows that it has important consequences for the residual entropy of ice. The problem has also been discussed by other authors (see, e.g., O'Konski, 1963).

It is evident that the "proton exchange" which occurs in connection with the hydrogen bonding is of great importance in the study of the tautomeric forms of the molecules involved, and that this exchange may be the main mechanism for approaching the so-called "tautomeric equilibrium." When L. Knorr in 1911 succeeded in separating the keto and enol forms of ethyl acetoacetate from each other, it appeared that, at room temperature, it would take weeks until the pure forms had reached the tautomeric equilibrium again. This time may be shortened, if there are additional protons available. However, if proton exchange and shifts between different tautomeric forms seem to be general features of systems having hydrogen bonds, one may wonder what precautions have to be taken to preserve a "proton code" in a stationary form for thousands of years, as in the DNA molecule in the Watson–Crick model. Apparently nature has succeeded very well.

Actual measurements of the tautomeric equilibrium constants for some of the pyrimidine bases have further given values in the range 10^{-3} to 10^{-5} (Kenner *et al.*, 1955; Katrizky and Waring, 1962, 1963) which indicate that the tautomeric forms are not as "rare" as Watson and Crick anticipated. This gives rise to interesting problems which will be discussed later in some greater detail.

B. Quantum Theory of the Hydrogen Bond: Tunnel Effect

The question of the motion of a proton in a hydrogen bond is, in a first rough approximation, a one-particle problem involving a fixed outer potential. Each electron pair attracts the proton, and this attraction may be represented by a single potential well, having a minimum corresponding to the classical equilibrium position. Since there are two electron pairs involved, the total effect is represented by a "double-well potential," of the form shown in Fig. 32. The double-well potential has two minima which are not necessarily of the same height and we will discuss both the symmetric and the asymmetric case. In classical physics, a particle would stay in one of the wells unless it has enough energy to pass the potential barrier separating the two minima. In quantum mechanics, however, the circumstances are slightly different depending on the fact that the proton is a "wave packet" which may also penetrate into classically forbidden regions. The phenomenon has an analog, e.g., in wave optics, where, in

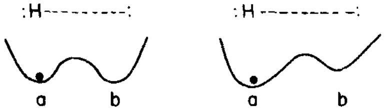


Fig. 32. Typical double-well potentials felt by a proton in a hydrogen bond.

connection with the total reflection of a light ray in a surface, the electromagnetic waves penetrate somewhat through the reflecting surface.

The very first study of the harmonic oscillator according to modern quantum theory showed that the particle involved could exist outside the classical "turning points" (see Fig. 33). This depends on the fact that the quantities kinetic energy and potential energy are not simultaneously measurable (see, e.g., Kramers, 1957, pp. 60, 61). The phenomenon implies that, if the potential energy curve shows two classically permitted intervals separated by a forbidden region, a quantum-mechanical particle may leak through the potential barrier from one "permitted" state to another. This phenomenon is known as the "tunnel effect" and is illustrated in Fig. 34.

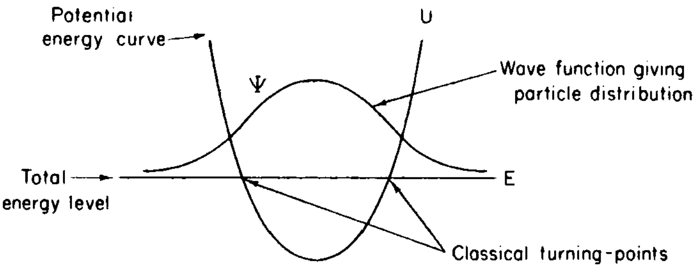


Fig. 33. The potential energy curve and the ground-state wave function of an harmonic oscillator.

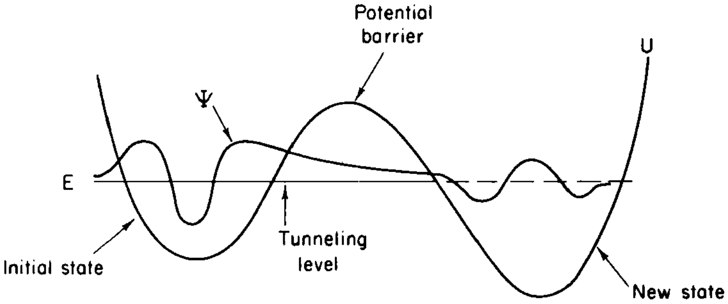


Fig. 34. Tunneling of a wave packet through a classically forbidden region.

The effect was first used by Gamow (1928a, b) and by Gurney and Condon (1929) independently of each other to explain the general phenomenon of *radioactivity*, which implies a transition from a bound state to a state in the continuum with a free particle emitted. The fact that the radioactive half-life times range from small fractions of seconds to thousands of years shows that tunneling probabilities can take values of all orders of magnitude. In molecular spectroscopy, the tunnel effect is quite well-known as causing the phenomenon of predissociation (see Herzberg, 1950; Landau and Lifshitz, 1958a). It is also of importance in the so-called ammonia clock. In solid state physics, the effect has been utilized for the technical construction of certain types of *semiconductors* known as tunneling diodes (Yajima and Esaki, 1958; Esaki, 1960). There seem also to be good reasons for believing that *corrosion* may be due to loss of order through particle tunneling.

In ordinary chemistry, the tunnel effect has so far been of smaller importance. In the theory of chemical kinetics, one would usually consider only processes which would have sufficient energy to take the components *above* the potential barrier between the two states involved, and the effect of tunneling is usually believed to be so small that it can be neglected. However, in certain biochemical processes where one has the effect of "biological amplification" it could very well happen that even the chemical effects of tunneling may show up, as we shall see below. For a study of the tunneling effect in ordinary chemical kinetics, we will refer to a paper by Hirschfelder and Wigner (1939).

Let us now return to the hydrogen bond and the double-well potential. It is evident that, if a proton can move even in a classically forbidden region, it may tunnel through the barrier from one equilibrium position to the other, and this property seems to be one of the characteristic features of the hydrogen bond.

The general properties of the hydrogen bond were discussed at the international symposium in Ljubljana in 1957 (Hadzi, 1959), and a great deal of interest was devoted to the question of the proton transfer in the bond (Coulson, 1959). It seems now as if the existence of the double-well potential has been experimentally rather well-established (Bell and Barrow, 1959; Haas and Hornig, 1959; Zimmermann, 1959, 1961; Joop and Zimmermann, 1962). For some recent developments in the theory of the hydrogen bond, we would like to refer to the 1957 symposium (Lipincott *et al.*, 1959) and to the works by Hofacker (1958), Fischer-Hjalmars and Grahn (1958), and Grahn (1959, 1961, 1962), Orgel (1959), and Huggins (1962).

Since we will later need a fair amount of details in our study of the hydrogen bonds of DNA, we will here give a brief survey of the quantum theory of a particle in a double-well potential.

1. Tunnel Effects in Rectangular Potential Barriers.

In order to study the basic features of the tunnel effect, one has to solve the Schrödinger equation

$$H_{op}\Psi = -\frac{h}{2\pi i}\frac{\partial\Psi}{\partial t}, \tag{IV.3}$$

where $H_{\mathrm{op}} = T + V$ is the Hamiltonian, T is the kinetic energy operator $p^2/2m = -(h^2/8\pi^2 m)\nabla^2$, and V is the potential energy. For a solution with a uniquely defined energy E, one has

$$\Psi = \Phi\, e^{-(2\pi i/h)Et}, \qquad H_{op}\Phi = E\Phi. \tag{IV.4}$$

For the sake of simplicity, we will study here only one-dimensional problems involving the coordinate x, so that $\Phi = \Phi(x)$ and $T = (h^2/8\pi^2 m)\partial^2/\partial x^2$. The main problem is now to solve the simplified Schrödinger equation:

$$-\frac{h^2}{8\pi^2 m}\frac{\partial^2\Phi}{\partial x^2} + V\Phi = E\Phi. \tag{IV.5}$$

One can get a very good idea of the nature of the tunnel effect by studying some simple model potentials (Fowler and Nordheim, 1928; Frank and Young, 1931), and we will limit ourselves here to rectangular potential barriers and start with the type given in Fig. 35. In the following, it is convenient to introduce the notations:

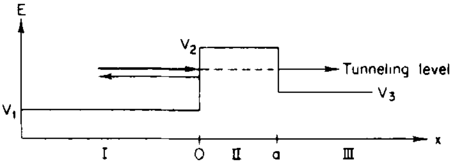


Fig. 35. Wave packet from the left penetrating a rectangular potential barrier.

$$p = \sqrt{2m(E - V)}, \qquad k = \frac{p}{\hbar} = \frac{2\pi}{h}\sqrt{2m(E - V)}, \tag{IV.6}$$

where p may be called the "classical momentum." Equation (IV.5) may now be rewritten in the form:

$$\frac{\partial^2\Phi}{\partial x^2} = -k^2\Phi. \tag{IV.7}$$

It is clear that, for the simple rectangular potential given above, it is easy to solve the Schrödinger equation, and one obtains the general solution

$$\Phi = A\, e^{ikx} + B\, e^{-ikx}. \tag{IV.8}$$

Let us now consider the special case when $V_3 < E < V_2$, i.e., the total energy of the particle is below the barrier but above both potential minima. For the three regions, one can now write the general solution in the form:

$$\begin{aligned} &(\mathrm{I})\ \Phi = A_1 e^{ik_1x} + B_1 e^{-ik_1x}, && k_1 = \frac{2\pi}{h}\sqrt{2m(E - V_1)}; \\ &(\mathrm{II})\ \Phi = A_2 e^{+|k_2|x} + B_2 e^{-|k_2|x}, && |k_2| = \frac{2\pi}{h}\sqrt{2m(V_2 - E)}; \\ &(\mathrm{III})\ \Phi = A_3 e^{ik_3x} + B_3 e^{-ik_3x}, && k_3 = \frac{2\pi}{h}\sqrt{2m(E - V_3)}. \end{aligned} \tag{IV.9}$$

These formulas will correspond to one and the same solution only if Φ and Φ' are continuous at the boundaries $x = 0$ and $x = a$, and this determines the connection between the coefficients. At $x = 0$ one has

$$\begin{cases} A_1 + B_1 = A_2 + B_2, \\ A_1 - B_1 = -i\dfrac{|k_2|}{k_1}(A_2 - B_2), \end{cases} \tag{IV.10}$$

and at $x =$ a one obtains similarly

$$\begin{cases} A_3 e^{ik_3a} + B_3 e^{-ik_3a} = A_2 e^{K} + B_2 e^{-K}, \\ A_3 e^{ik_3a} - B_3 e^{-ik_3a} = -i\dfrac{|k_2|}{k_3}(A_2 e^{K} - B_2 e^{-K}), \end{cases} \tag{IV.11}$$

where we have introduced the notation

$$K = a|k_2| = \frac{2\pi a}{h}\sqrt{2m(V_2 - E)}. \tag{IV.12}$$

Let us now study the reflection and transmission of the matter waves with respect to the barrier. The quantities

$$e^{(2\pi i/h)(px - Et)}$$

and

$$e^{-(2\pi i/h)(px - Et)}$$

correspond to waves which travel to the right and to the left, respectively, on the x axis. We will now consider a situation characterized by the fact that there is an incident wave coming in region I, so that $A_1 \neq 0$, but no incident wave in region III, so that $B_3 = 0$. The incident wave with the intensity $|A_1|^2$ may very well be reflected at the boundaries of the barrier, but it could happen that a proton penetrates the barrier and appears in region III as a right-going wave with the intensity $|A_3|^2$. The quotient

$$g = |A_3|^2/|A_1|^2 \tag{IV.13}$$

is called the "transmission coefficient" of the barrier. Putting $B_3 = 0$ in the connection formulas (IV.10) and (IV.11), and eliminating B_1, A_2, and B_2, one obtains directly

$$\frac{A_1}{A_3} e^{-ik_3 a} = \frac{1}{4}\left(1 + \frac{k_3}{k_1}\right)(e^K + e^{-K}) + \frac{i}{4}\left(\frac{|k_2|}{k_1} - \frac{k_3}{|k_2|}\right)(e^K - e^{-K}). \tag{IV.14}$$

For the inverse of the transmission coefficient, this gives

$$\boxed{g^{-1} = \frac{1}{4}\left(1 + \frac{k_3}{k_1}\right)^2 + \frac{1}{4}\frac{(k_1^2 + |k_2|^2)(k_3^2 + |k_2|^2)}{k_1^2|k_2|^2}\sinh^2 K} \tag{IV.15}$$

which shows the existence of the "tunnel effect." We note the complex character of the wave function involved, and the interesting problem of the relative phases in the three regions. Since $\sinh K = \frac{1}{2}(e^K - e^{-K})$, the quantity g^{-1} goes up exponentially as e^{2K}, and this implies that the transmission coefficient decreases as e^{-2K}:

$$g \approx e^{-2K}. \tag{IV.16}$$

This means that the effect goes down exponentially with the width of the barrier and with the square root of the energy distance to the top. This explains the enormous sensitivity of the entire tunnel effect and why tunneling times easily may vary over all orders of magnitude.

If we now change the left-hand side of region I and form a closed well, the particle is going to oscillate in this well. If it hits the barrier ν_1 times per second, the probability per second for tunneling to the right would be

$$S = \nu_1 g. \tag{IV.16a}$$

This is a formula which we will frequently use in the following.

We will now consider the problem from another point of view. For this purpose, we will introduce a rectangular double-well potential of the type indicated in Fig. 36 and ask for the stationary states of the particle. We will assume that the barriers at $x = -b_1$ and $x = a + b_3$ are infinitely high, so that one has the boundary conditions $\Phi = 0$ at these points.

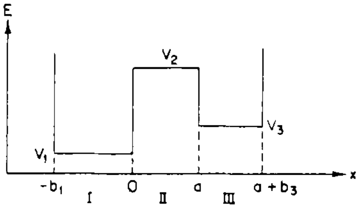


Fig. 36. Rectangular double-well potential.

Let us first consider the energy in the interval $V_1 < E < V_3$, in which case only regions I and II are practically involved. Putting $\Phi = 0$ at the point $x = -b_1$ gives the relation:

$$\cot k_1 b_1 = +i\,\frac{A_1 - B_1}{A_1 + B_1}. \tag{IV.17}$$

In region II, we will request that the solution decreases exponentially, which leads to the condition $A_2 = 0$. Combination of (IV.10) and (IV.17) leads then to the "quantum condition":

$$\cot k_1 b_1 = -\frac{|k_2|}{k_1}, \tag{IV.18}$$

which determines the energy levels in the range $V_1 < E < V_3$.

Next we will consider the "tunneling" interval $V_2 < E < V_3$, in which case one has an oscillatory solution also in region III. Application of the strict boundary condition $\Phi = 0$ at the point $x = a + b_3$, gives immediately the relation:

$$\cot k_3 b_3 = -i\,\frac{A_3 e^{ik_3 a} - B_3 e^{-ik_3 a}}{A_3 e^{ik_3 a} + B_3 e^{-ik_3 a}}, \tag{IV.19}$$

and combination of (IV.10), (IV.11), (IV.17), and (IV.19) will render the "quantum conditions." Combining (IV.10) and (IV.17) we first obtain

$$\frac{B_2}{A_2} = \frac{1 - \dfrac{k_1}{|k_2|}\cot k_1 b_1}{1 + \dfrac{k_1}{|k_2|}\cot k_1 b_1}, \tag{IV.20}$$

and combining (IV.11) and (IV.19) we get further

$$\frac{B_2}{A_2}e^{-2K}=\frac{1+\dfrac{k_3}{|k_2|}\cot k_3b_3}{1-\dfrac{k_3}{|k_2|}\cot k_3b_3}. \tag{IV.21}$$

For the interval $V_2 < E < E_3$, the quantum conditions may hence be written in the form:

$$\boxed{\begin{aligned}&\left(1+\frac{k_1}{|k_2|}\cot k_1b_1\right)\left(1+\frac{k_3}{|k_2|}\cot k_3b_3\right)=\\ &\qquad e^{-2K}\left(1-\frac{k_1}{|k_2|}\cot k_1b_1\right)\left(1-\frac{k_3}{|k_2|}\cot k_3b_3\right)\end{aligned}} \tag{IV.22}$$

which shows a nice symmetry with respect to the regions I and III.

The energy interval $E > V_2$ is treated analogously and, in this way, one obtains a complete knowledge of the stationary states of the double-well potential treated as a whole.

By means of the stationary states, one can now also study the time-dependent problems. If the initial state at $t = 0$ is given by the wave function $\Psi = \Psi_0$, the solution to the Schrödinger equation (IV.3) may be written in the form:

$$\Psi(x, t)=\sum_k e^{-2\pi i/h\cdot E_k\cdot t}\Phi_k(x)\langle\Phi_k\,|\,\Psi_0\rangle, \tag{IV.23}$$

which is known as the "expansion in the stationary states." Here we have used the notation $\langle\Phi_k|\Psi_0\rangle = \int \Phi_k{}^*\Psi_0(dx) = C_k$. Forming the probability density $|\Psi|^2$, one obtains

$$|\Psi(x, t)|^2=\sum_{kl} e^{2\pi i\nu_{kl}t}C_k^*C_l\Phi_k^*(x)\Phi_l(x), \tag{IV.24}$$

where $\nu_{kl} = (E_k - E_l)/h$ is the Bohr frequency for the energy levels involved. Depending on the fact that (IV.23) represents a coherent superposition of waves, the probability density is going to oscillate and, for the energy levels in the range $V_3 < E < V_2$, these oscillations will be extended over the region of the products $\Phi_k{}^*\Phi_l$, i.e., over the entire double well. From these results, one can easily derive the tunneling time and the transmission coefficients (Löwdin, 1964). It is clear that this method leads to more accurate results than the use of the formula (IV.16a), but the results are otherwise quite analogous.

2. *Tunnel Effect in an Arbitrary Barrier Treated by the JWKB Method*

In order to treat the tunnel effect in a potential barrier of a more arbitrary shape in a first approximation, it is feasible to use the so-called JWKB method (Jeffreys, 1923; Wentzel, 1926; Kramers, 1926; Brillouin, 1926a, b; see also Mott and Sneddon, 1948). Let us again start from the one-dimensional Schrödinger equation:

$$-\frac{h^2}{8\pi^2 m}\frac{\partial^2\Phi}{\partial x^2} + V(x)\Phi = E\Phi \tag{IV.25}$$

and introduce the functions

$$p(x) = \sqrt{2m\{E - V(x)\}}$$

and

$$k(x) = \frac{2\pi}{h}\,p(x). \tag{IV.26}$$

According to the JWKB scheme, an approximate solution may now be presented in the form:

$$\Phi \approx Ak^{-1/2}\exp\left(i\int^x k\,dx\right) + Bk^{-1/2}\exp\left(-i\int^x k\,dx\right). \tag{IV.27}$$

A point where $k = 0$ will be called a "turning point," and such a point is usually situated between a classically permitted and a classically forbidden region. The JWKB solution has oscillatory behavior in the "permitted" region and exponential behavior in the "forbidden" region [cf. (IV.9)].

Of fundamental importance is the Jeffreys–Kramers formula which connects the solutions on both sides of a turning point $x = x_1$ or $x = x_3$ (see Fig. 37). For real functions, it takes the form:

$$c_1 k^{-1/2}\sin\left\{\int_x^{x_1} k\,dx + \frac{\pi}{4} + \gamma_1\right\} \leftrightarrow$$

$$c_1|k|^{-1/2}\left\{\tfrac{1}{2}\cos\gamma_1\exp\left(-\int_{x_1}^x |k|\,dx\right) + \sin\gamma_1\exp\left(+\int_{x_1}^x |k|\,dx\right)\right\}, \tag{IV.28}$$

$$c_3|k|^{-1/2}\left\{\tfrac{1}{2}\cos\gamma_3\exp\left(-\int_x^{x_3} |k|\,dx\right) + \sin\gamma_3\exp\left(+\int_x^{x_3} |k|\,dx\right)\right\} \leftrightarrow$$

$$c_3 k^{-1/2}\sin\left\{\int_{x_3}^x k\,dx + \frac{\pi}{4} + \gamma_3\right\}. \tag{IV.29}$$

Fig. 37. The two types of turning points in the JWKB method.

For complex functions, one obtains instead:

$$\alpha_1 k^{-1/2}\exp\left[-i\left(\int_x^{x_1} k dx+\frac{\pi}{4}\right)\right]+\beta_1 k^{-1/2}\exp\left[+i\left(\int_x^{x_1} k dx+\frac{\pi}{4}\right)\right]\leftrightarrow \tag{IV.30}$$
$$\alpha_2|k|^{-1/2}\exp\left(+\int_{x1}^{x}|k|\,dx\right)+\beta_2|k|^{-1/2}\exp\left(-\int_{x_1}^{x}|k|\,dx\right),$$

with the connection formulas:

$$\alpha_2=\alpha_1+\beta_1$$

and

$$\beta_2=-\tfrac{1}{2}i(\alpha_1-\beta_1). \tag{IV.31}$$

Similarly one has around the other turning point:

$$\alpha_3 k^{-1/2}\exp\left[i\left(\int_{x_3}^{x} k dx+\frac{\pi}{4}\right)\right]+\beta_3 k^{-1/2}\exp\left[-i\left(\int_{x_3}^{x} k dx+\frac{\pi}{4}\right)\right]\leftrightarrow \tag{IV.32}$$
$$\alpha_2'|k|^{-1/2}\exp\left(+\int_{x}^{x_3}|k|\,dx\right)+\beta_2'|k|^{-1/2}\exp\left(-\int_{x}^{x_3}|k|\,dx\right),$$

with the connection formulas:

$$\alpha_2'=\alpha_3+\beta_3$$

and

$$\beta_2'=\tfrac{1}{2}i(\alpha_3-\beta_3). \tag{IV.33}$$

We will now use these formulas to study the quantum mechanics of a particle in a double-well potential, where the barrier has two turning points (see Fig. 38). As before, we will treat the tunnel effect by

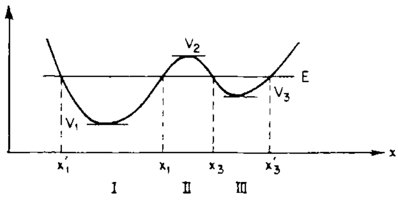


Fig. 38. Typical double-well potential.

considering both a single incident wave from the left and the more general problem of the stationary states.

a. Single Incident Wave. In order to derive a formula for the transmission coefficient in general, we will consider a single incident wave hitting a barrier with two turning points (see Fig. 39). We can proceed exactly as we did before in the case of a rectangular barrier. In region I there is an incident wave

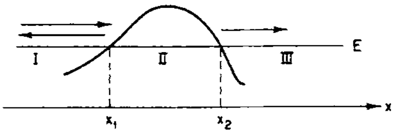


Fig. 39. Wave packet from the left penetrating a potential barrier.

$$\alpha_1 k^{-1/2} \exp\left[i\left(\int_{x_1}^{x} k\,dx + \frac{\pi}{4}\right)\right] \tag{IV.34}$$

which is partially transmitted into region III in the form of a right-going wave

$$\alpha_3 k^{-1/2} \exp\left[i\left(\int_{x_3}^{x} k\,dx + \frac{\pi}{4}\right)\right] \tag{IV.35}$$

whereas $\beta_3 = 0$. The somewhat modified transmission coefficient is now defined by the fraction $g = |\alpha_3|^2/|\alpha_1|^2$. Introducing the symbol

$$\boxed{K = \int_{x_1}^{x_3} |k|\,dx = \frac{2\pi}{h}\int_{x_1}^{x_3} \sqrt{2m[V(x) - E]}\,dx} \tag{IV.36}$$

and using the connection formulas (IV.31) and (IV.35), we obtain:

$$\begin{cases} \alpha_2' = \alpha_3, & \beta_2' = \frac{1}{2}i\alpha_3, \\ \alpha_2 = \beta_2'\,e^{-K} & \beta_2 = \alpha_2'\,e^{+K}, \\ \alpha_2 = \alpha_1 + \beta_1, & \beta_2 = -\frac{1}{2}i(\alpha_1 - \beta_1). \end{cases} \tag{IV.37}$$

This gives immediately the relations

$$\frac{\alpha_3}{\alpha_1} = -2i(\tfrac{1}{2}\,e^{-K} + 2\,e^{+K})^{-1}$$

and

$$\boxed{g = \frac{|\alpha_3|^2}{|\alpha_1|^2} = \frac{e^{-2K}}{(1 + \frac{1}{4}e^{-2K})^2} \approx e^{-2K}} \tag{IV.38}$$

and one can hence use formula (IV.16) for the transmission coefficient, provided K is defined by (IV.36).

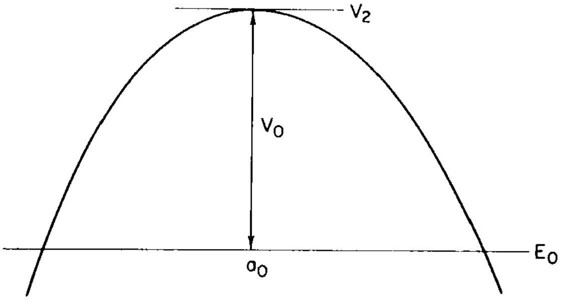


Fig. 40. Parabolic potential barrier.

If the barrier is parabolic at its top (see Fig. 40), one can easily evaluate the integral in (IV.36):

$$K_0 = \frac{\pi}{4} \cdot \frac{2\pi}{h} a_0 \sqrt{2m(V_2 - E_0)}. \tag{IV.39}$$

The factor $\pi/4$ comes here from the shape of the hill. As E approaches the top of the hill the tunneling becomes easier, since both the height and the width are decreasing to zero. If one measures the energy from the top in fractions κ of V_0, so that $E = V_2 - \kappa V_0$, one obtains

$$K = \kappa K_0, \tag{IV.40}$$

which formula is often useful.

b. Stationary States of the Double-Well Potential. One finds the stationary states of a double-well potential by solving the Schrödinger equation (IV.25). For a symmetric double well with $V_1 = V_3$, the solutions are necessarily symmetric with respect to inversion and are either "gerade" g or "ungerade" u. In both cases, the probability distribution is 50–50 over the two wells. For an asymmetric potential the circumstances are somewhat more complicated, and if $V_1 < E < V_3$ the particle will be essentially localized in the left-hand well. For $V_3 < E < V_2$ it will tunnel through the barrier and for $E > V_2$ it will pass over it.

Let us now calculate the energy eigenvalues and the stationary states of a double well by means of the JWKB method, and let us start by considering the case $V_1 < E < V_3$ (illustrated in Fig. 41). In this case, there are two turning points, $x = x_1'$ and $x = x_1$, involved in the problem, and the solutions are characterized by the fact that, outside these turning points, they should be exponentially decreasing. For $x = x_1$ one obtains, for instance, according to the connection formula (IV.28):

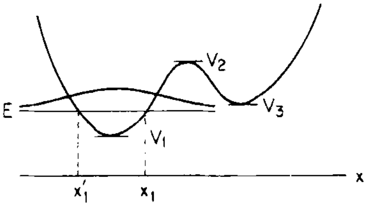


Fig. 41. Stationary state of a double-well potential having an energy below the lowest tunneling level.

$$c_1 k^{-1/2} \sin\left\{\int_x^{x_1} k\,dx + \frac{\pi}{4}\right\} \leftarrow c_1|k|^{-1/2}\tfrac{1}{2}\exp\left(-\int_{x_1}^{x}|k|\,dx\right), \qquad \text{(IV.41)}$$

and a similar relation for the other turning point:

$$-c_1|k|^{-1/2}\tfrac{1}{2}\exp\left(-\int_x^{x'_1}|k|\,dx\right) \rightarrow c_1 k^{-1/2}\sin\left\{-\int_{x'_1}^{x} k\,dx - \frac{\pi}{4}\right\}, \qquad \text{(IV.42)}$$

where we have enclosed an extra factor (-1). Since the expressions in the middle are identical, one obtains:

$$\int_x^{x_1} k\,dx + \frac{\pi}{4} = -\int_{x'_1}^{x} k\,dx - \frac{\pi}{4} + (n+1)\pi$$

$$\boxed{\int_{x'_1}^{x_1} k\,dx = (n + \tfrac{1}{2})\pi} \qquad \text{(IV.43)}$$

which is analogous to the ordinary eigenvalue relation in the JWKB method. [Compare also the exact expression, given by Milne (1930).]

Let us now consider the tunneling interval $V_3 < E < V_2$. In this case, there are four turning points, x_1', x_1, x_2, and x_2', and the solutions should be such that outside x_1' and x_2' they should be exponentially decreasing. Let us denote the phase constants in regions I and III by γ_1 and γ_3, respectively. Comparing the two exponential solutions in region II by means of (IV.28) and (IV.29), we obtain:

$$\begin{cases} \tfrac{1}{2}c_1 \cos\gamma_1 = c_3 \sin\gamma_3 \cdot e^{+K}, \\ c_1 \sin\gamma_1 = \tfrac{1}{2}c_3 \cos\gamma_3 \cdot e^{-K}, \end{cases} \tag{IV.44}$$

and further $tg\gamma_1 \cdot tg\gamma_3 = \frac{1}{4}e^{-2K}$. According to (IV.42), one has the connection formula:

$$-c_1|k|^{-1/2}\tfrac{1}{2}\exp\left(-\int_{x_1}^{x'_1}|k|\,dx\right) \rightarrow \\ c_1 k^{-1/2}\sin\left\{\int_x^{x_1} kdx + \int_{x_1}^{x'_1} kdx - \frac{\pi}{4}\right\}$$

which gives the phase $\gamma_1 = -[\int_{x'_1}^{x_1} kdx + (\pi/2)]$. A similar relation holds for $\gamma_3 = -[\int_{x_3}^{x'_3} kdx + (\pi/2)]$. Together with (IV.44), this gives the "quantum condition"

$$\boxed{tg\left(\int_{x'_1}^{x_1} kdx + \frac{\pi}{2}\right) \times tg\left(\int_{x_3}^{x'_3} kdx + \frac{\pi}{2}\right) = \tfrac{1}{4}\,e^{-2K}.} \tag{IV.45}$$

3. *Temperature Dependence of Tunnel Effect*

In order to discuss the influence of the temperature on the tunnel effect, we will consider a particle with mass m in a double-well potential and assume that it is in a weak thermal interaction with an environment having the temperature T. More specifically, one may think about a proton in a hydrogen bond in a particular base pair in a DNA molecule in its cellular environment, but we will here try to give the discussion a more general character. As the time $t = 0$, we will further assume that the particle is entirely in the deepest well and then study the approach to thermal equilibrium.

The final thermal equilibrium is a Boltzmann distribution over the energy levels E_0, E_1, E_2, ... of the double-well potential, characterized by the population factors

$$e^{-E_n/kT}, \tag{IV.46}$$

where $k = 8.6167 \times 10^{-5}$ eV/degree and T is the absolute temperature. Denoting the degeneracy of the level E_n by ω_n, and introducing the partition sum Z and the free energy F:

$$Z = \sum_n \omega_n \, e^{-E_n/kT} = Tr\{e^{-H_{op}/kT}\}, \qquad \text{(IV.47)}$$
$$F = -kT \log Z,$$

one obtains the probability P_n for finding the particle in the specific state E_n:

$$P_n = \omega_n \, e^{-E_n/kT} \cdot Z^{-1}$$
$$= \omega_n \, e^{-(E_n - F)/kT}. \qquad \text{(IV.48)}$$

In the following, we will be mainly interested in the relative distribution of the particle population, so we will essentially use only the Boltzmann factor (IV.46). The problem is now to calculate the distribution not only over the energy levels but also over the two wells.

For this purpose, we will consider an asymmetric double-well potential under the assumption that the particle is initially entirely to the left (see Fig. 42). It is convenient to distinguish between two cases, namely fast and slow tunneling. In the slow tunneling case, the penetration of the barrier is so small that an approximate thermal equilibrium will first be formed in the left-hand well and, during the course of time, the particle will then pass the barrier so that the system may approach the final equilibrium. In the fast tunneling case, the penetration of the barrier is instead so high that the particle immediately passes the barrier and the system goes directly to the final thermal equilibrium over the entire double well. The distinction depends obviously on how the "tunneling times" are related to the time intervals necessary to reach thermal equilibrium. Before we can proceed, it is necessary to discuss the concept of tunneling time in greater detail.

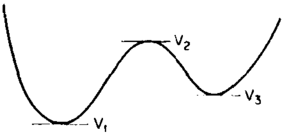


Fig. 42. Typical double-well potential.

a. Tunneling out of a Single Well. Let us first consider a double-well potential, where the right-hand well has been smeared out so that there is no "back-tunneling" (see Fig. 43). For the sake of simplicity, we will assume that the deepest well is essentially parabolic and characterized by the potential $\alpha_1(x - x_1)^2/2$. A classical particle in such a well will be a harmonic oscillator with the frequency $\nu_1 = (\alpha_1/4\pi^2 m)^{1/2}$, which is a quantity proportional to $m^{-1/2}$. Elementary quantum mechanics tells us

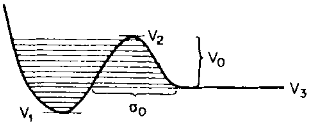


Fig. 43. Single-well potential with a large number of tunneling levels.

now that the "wave packet" representing our particle moves in such a way that its center oscillates like a classical particle, and this means that, irrespective of the energy level, the wave packet is going to hit the barrier ν_1 times per second. At each hit, there is the transmission probability g given by formula (IV.38), which means that

$$c_1 = \nu_1 g \tag{IV.49}$$

is the fraction of a particle which is going to tunnel per second at a given energy level provided that a "full" particle would hit the barrier every time. The quantity c_1 will be denoted as the "tunneling rate," whereas the quantity

$$\tau_1 = 1/c_1 = 1/\nu_1 g \tag{IV.50}$$

will be called the "tunneling time."

In the following, we will assume that the transmission coefficient g is very small, so that $g \ll 1$. Even in this case, one has to consider the change of the population in the left-hand well. At every hit, the population in a particular energy level will decrease by the factor $(1 - g)$ and, after p hits, the decrease will be described by the factor $(1 - g)^p$. If the energy level population is originally a_1, the population after the time t sec will hence be

$$n_1(t) = a_1(1 - g)^{\nu_1 t}, \tag{IV.51}$$

which is a step function. However, if the transmission coefficient g is sufficiently small, one should observe that the quantity $(1 - 1/g^{-1})^{g^{-1}}$ may be approximated by the number e^{-1}, which leads to the formula:

$$\begin{aligned} n_1(t) &= a_1\{(1 - 1/g^{-1})^{g^{-1}}\}^{g\nu_1 t} \\ &\approx a_1 e^{-c_1 t}. \end{aligned} \tag{IV.52}$$

This means that the step function may be replaced by a function having a continuous derivative and satisfying the differential equation:

$$\frac{dn_1}{dt} = -c_1 n_1. \tag{IV.53}$$

The use of this relation will greatly simplify the tunneling problem in the following. The part of the population which has passed the barrier during the time t is hence:

$$n_3(t) = a_1(1 - e^{-t/\tau_1}), \tag{IV.54}$$

and for $t \ll \tau_1$ one may hence use the linear relation $n_3(t) \approx a_1 c_1 t$. The shape of the function (IV.54) is illustrated in Fig. 44. The "tunneling time" τ defined by (IV.50) is hence the time required for the original population to decrease by a factor e^{-1}.

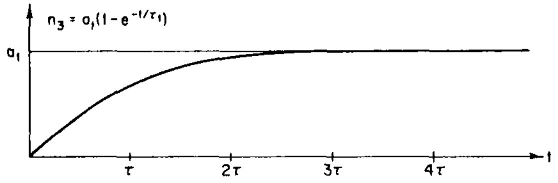


Fig. 44. Tunneling out of a single-well potential as a function of time.

Slow tunneling. We will now consider the case when the tunneling time τ is long in comparison with the time needed for the particle to reach a thermal equilibrium with the environment. In this case, there will first be established an approximate thermal equilibrium in the left-hand well characterized by the Boltzmann distribution:

$$P_n = P_1 e^{-(E_n - E_1)/kT}, \tag{IV.55}$$

where P_1 is the ground-level population of the left-hand well. In the following, we will use the notation $E_n = E = V_2 - \kappa V_0$, so that $\kappa = 0$ and $\kappa = 1$ correspond to the top and the bottom of the tunneling barrier, respectively. The transmission coefficient g may be found by a combination of formulas (IV.38) to (IV.40), which gives:

$$g = \exp\left(-\frac{\pi^2}{h}\kappa a_0 \sqrt{2mV_0}\right). \tag{IV.56}$$

For $t \ll \tau$, the population of the level $E = V_2 - \kappa V_0$ passing through the barrier is then given by the linear relation $n_3(t) \approx a_1 v_1 g t$ or:

$$n_3(t) = v_1 t P_1 \exp\left\{-(V_2 - V_1)/kT + \kappa\left(\frac{V_0}{kT} - \frac{\pi^2}{h} a_0 \sqrt{2mV_0}\right)\right\}. \quad \text{(IV.57)}$$

The easiness for tunneling increases towards the top of the barrier, whereas the Boltzmann distribution decreases, and the last factor reflects the competition between these two effects. For a special temperature T_0 called the "characteristic temperature" and defined by the relation

$$\frac{\pi^2}{h} a_0 \sqrt{2mV_0} = \frac{V_0}{kT_0}, \quad \text{(IV.58)}$$

the two effects cancel each other. This temperature may be used to characterize the essential features of the tunneling effect, and it is related to the properties of the barrier by the formula:

$$T_0 = \frac{h}{\pi^2 k} \frac{\sqrt{V_0/2m}}{a_0}. \quad \text{(IV.59)}$$

One can then conveniently write (IV.57) under the form:

$$n_3(t) = v_1 t P_1 \exp\left\{-\frac{V_2 - E_1}{kT} + \frac{\kappa V_0}{k}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right\}. \quad \text{(IV.60)}$$

This formula shows that for $T < T_0$ the tunneling is largest towards the bottom of the barrier ($\kappa = 1$) and decreases towards the top ($\kappa = 0$), whereas for $T > T_0$ the reverse is true. Here we will essentially consider only the former case.

Let us now evaluate the total tunneling out of the barrier during the time t, when $t \ll \tau$ and $T \ll T_0$. For the sake of simplicity, we will assume the existence of an average level spacing hv_1, so that $\Delta\kappa = hv_1/V_0$. This implies that the quantity (IV.60) for $\kappa = 1$, $1 - \Delta\kappa$, $1 - 2\Delta\kappa$, ... forms a geometrical progression with the quotient

$$\exp\left\{-\frac{hv_1}{k}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right\} \quad \text{(IV.61)}$$

and the approximate sum

$$N_3(t) = c_{1/\kappa=1} t \cdot P_1 e^{-(V_3 - V_1/kT)} \times \frac{1 - \exp\left\{-\frac{V_0}{k}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right\}}{1 - \exp\left\{-\frac{hv_1}{k}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right\}}. \quad \text{(IV.62)}$$

The essential conclusion is that $\log(N_3/t)$ is no longer proportional to $1/kT$ and that for large values of this quantity, i.e., small temperature T, one has

$$\log(N_3/P_1t) \approx \log c_{1/\kappa=1} - \frac{V_3 - V_1}{kT} + \cdots. \tag{IV.63}$$

An experimental study of this quantity should hence give essential information about the tunnel effect according to (IV.56).

Estimate of the oscillatory frequency ν_1. The Hamiltonian for the harmonic oscillator in the left-hand well is given by the expression:

$$H = \frac{p^2}{2m} + \frac{1}{2}\alpha_1(x - x_1)^2. \tag{IV.64}$$

Consequently, if $2b_1$ is the distance between the two classical turning points for the ground state, one has the relation $E_1 = \frac{1}{2}\alpha_1 b_1{}^2 = h\nu_1/2$. Since, further, $\alpha_1 = 4\pi^2 m\nu_1{}^2$, this gives

$$\nu_1 = \frac{h}{4\pi^2 m b_1^2}. \tag{IV.65}$$

A similar formula may be derived from Heisenberg's uncertainty relations by letting $\Delta x = b_1/\sqrt{2}$ be the uncertainty of the position of the particle.

If the exact form of the well is known, one can determine the oscillatory frequency ν_1 for each level by solving the Schrödinger equation. In this connection, one has to distinguish between the Bohr frequency E/h for the level and the "hit" frequency with respect to the barrier. This is clearly illustrated by the harmonic oscillator having the energy levels $(n + \frac{1}{2})h\nu_1$, where n is associated with the number of nodes of the wave function, whereas ν_1 is the hit frequency characteristic for the motion of the wave packet.

b. Tunneling in a Double Well. After these preliminary studies dealing with some fundamental concepts concerning the tunnel effect, we will now consider the tunneling in a double well in greater detail. Particular attention has to be devoted to the question of "back-tunneling" from the more shallow to the deeper well.

We will assume that initially at the time $t = 0$ the particle is entirely in the left-hand well having the part a_1 of its population in a specific energy level E with the transmission coefficient $g = g(E)$. Further, let $n_1 = n_1(t)$ and $n_3 = n_3(t)$ denote the parts of the population in this level at time t associated with the left-hand and right-hand wells, respectively. If there are no other perturbations, one has $n_1 + n_3 = a_1$, and the first question is how the population a_1 is divided between the two wells as a function of time.

The problem may be treated by solving the time-dependent Schrödinger equation for the motion of a wave packet. Here we will instead consider an assembly for which one may use the law of additivity of transition probabilities. The tunneling rates are given by the quantities

$$c_1 = v_1 g \quad \text{and} \quad c_3 = v_3 g, \tag{IV.66}$$

and by a generalization of (IV.53) one obtains a system of differential equations:

$$\begin{cases} \dfrac{dn_1}{dt} = -c_1 n_1 + c_3 n_3, \\[2ex] \dfrac{dn_3}{dt} = c_1 n_1 - c_3 n_3. \end{cases} \tag{IV.67}$$

The initial condition $n_1(0) = a_1$, $n_3(0) = 0$ corresponds to the solution:

$$\begin{cases} n_1(t) = \dfrac{c_3 a_1}{c_1 + c_3} + \dfrac{c_1 a_1}{c_1 + c_3} e^{-(c_1+c_3)t}, \\[2ex] n_3(t) = \dfrac{c_1 a_1}{c_1 + c_3} - \dfrac{c_1 a_1}{c_1 + c_3} e^{-(c_1+c_3)t}. \end{cases} \tag{IV.68}$$

The time $\tau_{1,3} = (c_1 + c_3)^{-1}$ will be denoted as the resulting "tunneling time" or relaxation time in the double-well potential. For the special case when $c_1 = c_3$, the character of the solutions are illustrated in Fig. 45.

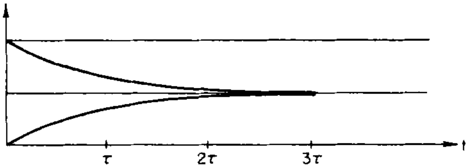


Fig. 45. Change of population in a tunneling level in a double-well potential as a function of time.

For $t \ll \tau_{1,3}$, the tunneling part of the population is given by the formula:

$$n_3 \approx a_1 c_1 t - a_1 c_1 (c_1 + c_3) \frac{t^2}{2} + \cdots. \tag{IV.69}$$

Here the linear term is associated with only left-right tunneling, whereas the quadratic term contains also the first effect of the "back-tunneling."

For $t \gg \tau_{1,3}$, the solutions have an entirely different character, since n_1 and n_3 quickly approach the constant values:

$$n_1 = \frac{\nu_3 a_1}{\nu_1 + \nu_3} \quad \text{and} \quad n_3 = \frac{\nu_1 a_1}{\nu_1 + \nu_3}, \tag{IV.70}$$

which satisfy the simple relation $n_1 : n_3 = \nu_3 : \nu_1$ or $n_1 \nu_1 = n_3 \nu_3$. This implies that for $t \gg \tau_{1,3}$ the distribution is essentially independent of the transmission coefficient g, i.e., of the shape of the barrier and of the particle mass m. In this case, the system is approaching a thermal equilibrium over the energy levels of the double well with the population of each level divided into the fractions $\nu_3 : \nu_1$ over the two wells.

This result is of particular importance in the case when the total barrier $(V_2 - V_1)$ is so high that the particle transfer above the barrier is negligible in comparison to the tunneling. As soon as $t \gg 1/g(\nu_1 + \nu_3)$, one can no longer distinguish from the populations how, for instance, different particle masses m have influenced the tunneling. This similarity between protons and deuterons, for example, for $t \gg \tau$ has to be remembered in interpreting the results of many experiments aimed at studying the tunneling phenomena in general (Pollard and Lemke, 1965; Löwdin, 1965).

In order to investigate the details of a tunneling process, it is hence necessary to go to the other extreme, namely $t \ll \tau$. In this case, the tunneling formula $n_3 = a_1 c_1 t$ for a double well is not yet influenced by the "back-tunneling," and formula (IV.57) is valid. In the case of *slow tunneling*, an approximate thermal equilibrium is first established in the *left-hand* well, and the total particle transmission through the barrier is then regulated by formulas (IV.62) and (IV.63). For intermediate time intervals, $t \approx \tau$, a special study has to be performed.

4. *Some Data Concerning Proton Tunneling in a Hydrogen Bond*

Let us now study the implications of the formulas in this section with respect to the proton tunneling in a hydrogen bond. Since we start out from quantum mechanics, it may be convenient to express the formulas first in atomic units ($e = 1$, $h/2\pi = 1$, $m_{el} = 1$) and later to convert to other units. In atomic units, the unit for length is 1 Bohr = 0.52917 Å, and the unit for energy is 1 Hartree = 27.2097 eV. The electron mass is $m = 1$, the proton mass is $m = 1840$, and the deuteron mass is $m = 3680$.

According to (IV.56), one obtains for the transmission coefficient in atomic units:

$$g = \exp\left(-\frac{\pi}{2}\kappa a_0\sqrt{2mV_0}\right). \tag{IV.71}$$

For practical applications, it is often more convenient to express the tunneling distance a_0 in angstrom units ($=10^{-8}$ cm) and the tunneling barrier V_0 in electron volts, which leads to the modified formula:

$$g = \exp(-0.80478\,\kappa a_0\sqrt{mV_0}). \tag{IV.72}$$

We recall that κ is a parameter used in defining the energy $E = V_2 - \kappa V_0$, so that $\kappa = 0$ and $\kappa = 1$ correspond to the top and the bottom of the tunneling barrier, respectively. For the electron, proton, and deuteron, one has consequently:

$$\begin{aligned} \text{electron: } g &= e^{-0.80478\,\kappa a_0\sqrt{V_0}} = 10^{-0.34951\,\kappa a_0\sqrt{V_0}}, \\ \text{proton: } g &= e^{-34.525\,\kappa a_0\sqrt{V_0}} = 10^{-14.994\,\kappa a_0\sqrt{V_0}}, \\ \text{deuteron: } g &= e^{-48.825\,\kappa a_0\sqrt{V_0}} = 10^{-21.213\,\kappa a_0\sqrt{V_0}}, \end{aligned} \tag{IV.73}$$

where a_0 is expressed in angstroms, V_0 in electron volts.

According to the data available in the literature, the lengths of the hydrogen bonds vary between 2.5 Å and 3 Å. For the sake of simplicity, we will here consider only the latter value and assume that the distance is equally divided by the equilibrium points, as illustrated in Fig. 46. The tunneling length a_0 varies between 0.8 Å and smaller values depending on the asymmetry between the two wells, and we will here consider the values 0.8 Å, 0.6 Å, and 0.4 Å as typical cases.

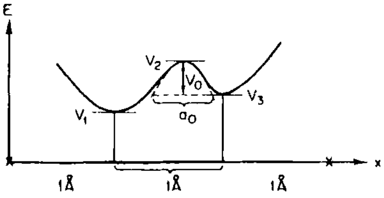


Fig. 46. Typical double-well potential felt by a proton in a hydrogen bond.

In order to determine the frequency ν_t for a proton, we will apply formula (IV.65) which takes the form:

$$\nu_t = \frac{1.0013}{b_1^2} \times 10^{12}\ \text{sec}, \tag{IV.74}$$

where $2b_l$ is the distance between the two turning points in angstrom units. Considering the bond distance and the shape of the well, the value $2b_l \approx 0.6$ Å seems reasonable and gives the frequency $\nu_l \approx 10^{13}$ sec^{-1} and the energy spacing 0.04 eV in good agreement with infrared data. The discussion in the following deals only with orders of magnitude, and we will hence rather arbitrarily put $\nu_l = 10^{13}$ sec^{-1} and $h\nu_l = 0.04135$ eV to have a uniquely defined reference point.

According to (IV.49) and (IV.73), the tunneling rate for a proton is now given by the expression:

$$c_l = \nu_l g = 10^{13-14.994\,\kappa a_0\sqrt{V_0}}. \tag{IV.75}$$

In Table II, the tunneling rates for the bottom of the barrier ($\kappa = 1$) and

TABLE II

TUNNELING RATES c_1 AND TUNNELING TIMES $\tau_1 = 1/c_1$ FOR THE ELECTRON, PROTON, AND DEUTERON AS FUNCTIONS OF THE PARAMETER $a_0\sqrt{V_0}$(a_0 IN Å, V_0 IN eV)[a]

	Electron	Proton		Deuteron	
$a_0\sqrt{V_0}$	$^{10}\log c_1$	$^{10}\log c_1$	τ_1	$^{10}\log c_1$	τ_1
0	14.6324	13.0	10^{-13} sec	12.8505	10^{-13} sec
0.1	14.5975	11.5	$10^{-11.5}$ sec	10.7292	10^{-11} sec
0.2	14.5625	10.0	10^{-10} sec	8.6079	10^{-9} sec
0.3	14.5276	8.5	$10^{-8.5}$ sec	6.4866	10^{-6} sec
0.4	14.4926	7.0	10^{-7} sec	4.3653	10^{-4} sec
0.5	14.4577	5.5	$10^{-5.5}$ sec	2.2440	10^{-2} sec
0.6	14.4227	4.0	10^{-4} sec	0.1227	1.3 sec
0.7	14.3878	2.5	10^{-2} sec	−1.9986	2 min
0.8	14.3528	1.0	10^{-1} sec	−4.1199	3.7 hrs
0.9	14.3178	−0.5	3 sec	−6.2412	20 days
1.0	14.2829	−2.0	2 min	−8.3625	8 years
1.1	14.2479	−3.5	53 min	−10.4838	10^3 years
1.2	14.2130	−5.0	1.16 days	−12.6051	1.3×10^5 years
1.3	14.1781	−6.5	1.22 months	−14.7264	1.7×10^7 years
1.4	14.1431	−8.0	3 years	−16.8477	2.2×10^9 years
1.5	14.1082	−9.5	100 years		
1.6	14.0732	−11.0	3000 years		
1.7	14.0383	−12.5	10^5 years		
1.8	14.0033	−14.0	3×10^6 years		
1.9	13.9684	−15.5	10^8 years		
2.0	13.9334	−17.0	3×10^9 years		
	$\tau_1 \approx 10^{-14}$sec				

[a] Electron: $\log c_1 = 14.6324 - 0.34951\, a_0\sqrt{V_0}$; proton: $\log c_1 = 13 - 15\, a_0\sqrt{V_0}$; deuteron: $\log c_1 = 12.8505 - 21.213\, a_0\sqrt{V_0}$.

the corresponding tunneling times τ are listed for $a_0\sqrt{V_0} = 0(0.1)2$ eV. It is remarkable that for this short range of the parameter $a_0\sqrt{V_0}$ the tunneling time τ varies between 10^{-13} sec and 10^{17} sec, which latter figure is of the same order as the age of the universe. For comparison, the tunneling data for the electron and the deuteron are also included; the corresponding values of v_l are by a classical analog chosen to be $v_l\sqrt{1840}$ and $v_l/\sqrt{2}$, respectively, whereas the transmission coefficients are given by (IV.73).

In Table III, the tunneling rate and the characteristic temperature T_0

TABLE III

Tunneling Rate c_1 and Characteristic Temperature T_0 for the Proton as Functions of V_0 (in eV) for the Tunneling Distances $a_0 = 0.8$Å, 0.6Å, and 0.4Å[a]

	$a_0 = 0.8$ Å		$a_0 = 0.6$ Å		$a_0 = 0.4$ Å	
V_0	$^{10}\log c_1$	T_0	$^{10}\log c_1$	T_0	$^{10}\log c_1$	T_0
0	13.0000	0	13.0000	0	13.0000	0
0.1	9.2052	132.88	10.1539	177.18	11.1026	265.77
0.2	7.6335	187.92	8.9751	250.57	10.3167	375.85
0.3	6.4274	230.16	8.0705	306.88	9.7137	460.32
0.4	5.4105	265.76	7.3079	354.35	9.2052	531.53
0.5	4.5147	297.13	6.6360	396.18	8.7573	594.27
0.6	3.7048	325.49	6.0286	433.99	8.3524	650.99
0.7	2.9601	351.57	5.4701	468.76	7.9800	703.15
0.8	2.2668	375.85	4.9501	501.13	7.6334	751.70
0.9	1.6158	398.65	4.4619	531.53	7.3079	797.30
1.0	1.0000	420.21	4.0000	560.28	7.0000	840.43
1.1	0.4144	440.72	3.5608	587.63	6.7072	881.45
1.2	0.8552-1	460.32	3.1414	613.76	6.4276	920.64
1.3	0.3176-1	479.12	2.7382	638.82	6.1588	958.24
1.4	0.8016-2	497.20	2.3512	662.93	5.9008	994.41
1.5	0.3036-2	514.65	1.9777	686.20	5.6518	1029.31
1.6	0.8212-3	531.53	1.6159	708.70	5.4106	1063.07
1.7	0.3544-3	547.89	1.2658	730.52	5.1772	1095.79
1.8	0.9008-4	563.78	0.9256	751.69	4.9504	1127.56
1.9	0.4592-4	579.22	0.5944	772.29	4.7296	1158.45
2.0	0.0296-4	594.27	0.2722	792.36	4.5148	1188.55
3.0	0.2148-8	727.83	0.4111-3	970.43	2.6074	1455.67
4.0	0.0000-11	840.43	0.0000-5	1120.56	1.0000	1680.86
		$420.21\sqrt{V_0}$		$560.28\sqrt{V_0}$		$840.43\sqrt{V_0}$

[a] $^{10}\log c_1 = 13 - 15a_0\sqrt{V_0}$; $T_0 = 336.17\sqrt{V_0}/a_0$.

for the proton are given as functions of V_0 for the three values $a_0 = 0.8$ Å, 0.6 Å, and 0.4 Å. For the calculation, one has used (IV.59) in the form:

$$T_0 = 336.17 \frac{\sqrt{V_0}}{a_0}. \qquad \text{(IV.76)}$$

The table among other things shows how essential small changes in a_0 may influence the tunneling rates, and this implies that in a more detailed study it is necessary to incorporate also the effect of the thermal vibrations of the molecules participating in the hydrogen bond under consideration.

TABLE IV

BOLTZMANN FACTORS $B = \exp\{-\Delta V/kT\}$ FOR $T = 273°\text{K}$, 310°K, AND 373°K AS FUNCTIONS OF ΔV(IN ev)

ΔV	$T = 273°\text{K}$, ${}^{10}\log B$	$T = 310°\text{K}$, ${}^{10}\log B$	$T = 373°\text{K}$, ${}^{10}\log B$
0	0	0	0
0.1	−1.846	−1.626	−1.351
0.2	−3.692	−3.252	−2.702
0.3	−5.538	−4.878	−4.053
0.4	−7.384	−6.504	−5.404
0.5	−9.230	−8.130	−6.755
0.6	−11.076	−9.756	−8.106
0.7	−12.922	−11.382	−9.457
0.8	−14.768	−13.008	−10.808
0.9	−16.614	−14.634	−12.159
1.0	−18.460	−16.260	−13.510
1.5	−27.690	−24.390	−20.265
2.0	−36.920	−32.520	−27.020
2.5	−46.150	−40.650	−33.775
3.0	−55.380	−48.780	−40.530
3.5	−64.610	−56.910	−47.285
4.0	−73.840	−65.040	−54.040
4.5	−83.070	−73.170	−60.795
5.0	−92.300	−81.300	−67.550
	$-18.46\Delta V$	$-16.26\Delta V$	$-13.51\Delta V$

Table IV gives the Boltzmann factor as a function of the potential energy difference ΔV in electron volts for some special temperatures. In this connection, it is convenient to write the Boltzmann factor in the form:

$$B = e^{-\Delta V/kT} = 10^{-K'' \cdot \Delta V}, \qquad \text{(IV.77)}$$

where $K'' = {}^{10}\log e/kT = 5040/T$. For the freezing and boiling points of water the values are $K'' = 18.46$ and $K'' = 13.51$, respectively, whereas for the body temperature of human beings ($T = 310°\mathrm{K}$) one has $K'' = 16.26$. This gives an idea of the K'' values which one meets in biological systems.

In the treatment of the double-well potential, two Boltzmann factors are of particular importance, namely $B_{1,2}$ associated with the total potential barrier $(V_2 - V_1)$ and $B_{1,3}$ associated with the energy difference $(V_3 - V_1)$ between the two minima. Since the fraction of the proton which is in the "tunneling region" is essentially proportional to the difference $(B_{1,3} - B_{1,2})$, it is clear that tunneling will be of importance in comparison to the transfer above the barrier if and only if $(B_{1,3} - B_{1,2}) \gg B_{1,2}$, which implies

$$B_{1,3} \gg \tfrac{1}{2} \cdot B_{1,2}. \tag{IV.78}$$

Table IV shows directly whether this is the case or not for a particular double well.

In the case of *slow tunneling*, the initial period of the tunneling process $(t \ll \tau)$ is regulated by the transmission formula (IV.63), which may be written:

$$N_3(t) = c_1 t \cdot P_1 B_{1,3} \times \frac{1 - \exp\left\{-\frac{V_0}{k}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right\}}{1 - \exp\left\{-\frac{h\nu_1}{k}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right\}}, \tag{IV.79}$$

where the last factor gives at least approximately the result of summing over all tunneling levels. For numerical purposes and the parameter values $T = 310°\mathrm{K}$ and $h\nu_1 = 0.04135$ eV, this factor may be expressed in the form

$$S = \frac{1 - 10^{-16.26 V_0 (1 - 310/T_0)}}{1 - 10^{-0.6723(1 - 310/T_0)}}. \tag{IV.80}$$

The second term in the numerator may often be neglected in comparison to the first.

In this discussion, we have not specified the three potential energy values V_1, V_2, and V_3 which characterize the shape of the double-well potential. It is immediately evident that the results are extremely sensitive with respect to small modifications of the potential energy curve, and that a change of only 0.1 eV in V_2 or V_3 changes the corresponding Boltzmann factor by a factor $10^{-1.6}$. A high degree of accuracy in the energy calculations is then desired.

In conclusion, it should be observed that the hydrogen bond is here treated by considering a one-dimensional model, whereas in reality the problem has three-dimensional character. This is going to influence the transmission coefficients and the tunneling times, so this problem has to be investigated in full detail.

Study of double-well potentials by means of superposition of model potentials. In the previous part of this section, we have simply characterized a double-well potential by means of its three extreme values V_1, V_2, and V_3 and the tunneling distance a_0 and assumed that the maxima and minima have roughly a parabolic shape. It is clear that this is a rather crude approximation, and one can now make a first refinement by describing the double well more consistently as a superposition of such model potentials as the Morse potential and the Born–Mayer potential. The Morse potential may be used to describe the attraction between a proton and a neutral molecule, whereas the Born–Mayer potential is characteristic for the attraction between a proton and a negative ion. In studying the potential along a specific line, it is not necessary to consider any details as to the angular dependence.

Let us start with the simplest case which consists of an "outside" proton which is attracted to two identical neutral molecules, so that the entire system forms an ion with a single positive charge. In such a case, the proton is influenced by a potential which may be approximately considered as the superposition of two Morse potentials.

In molecular theory, the Morse potential has proven to be a most valuable tool. It is characterized by three parameters D, a, and R_0:

$$\begin{aligned} V(R) &= D\{e^{-2a(R-R_0)} - 2e^{-a(R-R_0)}\} \\ &= -D + D\{e^{-a(R-R_0)} - 1\}^2, \end{aligned} \tag{IV.81}$$

and it describes a single well with the minimum $-D$ for $R = R_0$, which for $R \approx R_0$ has the form $V(R) \approx -D + Da^2(R - R_0)^2$ corresponding to a harmonic oscillator with the classical frequency ν determined by the relation $Da^2 = 2\pi^2 m\nu^2$. This gives the connection

$$a = 2\pi\nu\sqrt{m/2D}, \tag{IV.82}$$

where ν is the "stretching frequency" associated with the minimum. As Morse showed, the Schrödinger equation is exactly soluble for this potential and has the energy levels:

$$E_n = -D + (n + \tfrac{1}{2})h\nu - \left(\frac{h^2\nu^2}{4D}\right)(n + \tfrac{1}{2})^{1/2}. \tag{IV.83}$$

The number of discrete levels is finite and regulated by the inequality

$$(n + \tfrac{1}{2}) \leq 2D/h\nu.$$

In order to work with the Morse potential, it is convenient to bring it to a form which is independent of the parameters. Instead of a, it is feasible to introduce a distance $d = a^{-1}$ and the variable $x = (R - R_0)/d$ which gives the expression

$$\frac{V}{D} = e^{-2x} - e^{-x}. \tag{IV.84}$$

In the treatment of the Morse potential, the distance d is hence the convenient unit of length.

Let us now construct the superposition of two equivalent single wells having their original minima a distance $x_0 d$ apart, so that

$$W/D = \{e^{-2x} - 2e^{-x}\} + \{e^{-2(x_0 - x)} - 2e^{-(x_0 - x)}\}, \tag{IV.85}$$

which means that the potentials are turned opposite to each other. The curve W/D is symmetric around the point $x = x_0/2$ and hence has an extreme value $W/D = 2(\alpha_0 - 1)^2 - 2$ at this point. In order to describe the extreme values, we have here introduced the parameter

$$\alpha_0 = e^{-x_0/2}. \tag{IV.86}$$

It appears that for $\alpha_0{}^2 > \frac{1}{4}$ or $x_0 < 0.693149$ the curve has a *single broad minimum* situated in the point $x = x_0/2$. For $\alpha_0{}^2 < \frac{1}{4}$ or $x_0 > 0.693149$ the curve has two minima and a maximum and is hence really a *double-well potential.* The extreme values are characterized by the following formulas:

$$\begin{array}{ll} \text{Maximum:} & \text{Minima:} \\ x = x_0/2, & x = -{}^{e}\log(\tfrac{1}{2} \pm \sqrt{\tfrac{1}{4} - \alpha_0^2}), \\ W/D = 2(\alpha_0 - 1)^2 - 2, & W/D = -1 - 2\alpha_0^2. \end{array} \tag{IV.87}$$

The quantity W/D as a function of x for certain values of x_0 is graphically illustrated in Fig. 47. Forming the difference between the extreme values, one obtains the potential barrier

$$V_0/D = (2\alpha_0 - 1)^2. \tag{IV.88}$$

One can then derive the potential barrier as a function of the distance between the two minima (see Fig. 48). A study of the experimental proton

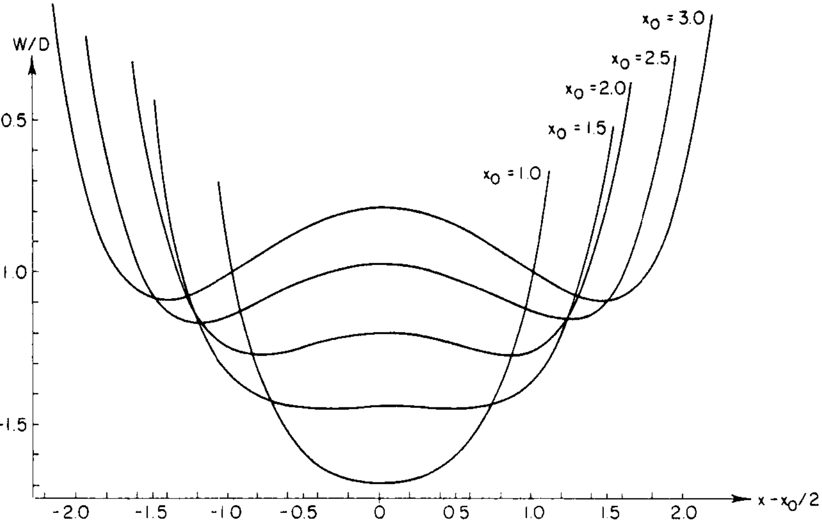


Fig. 47. Symmetric double-well potentials obtained by superposition of two Morse potentials.

data available in the symmetric case reveals that the length of the hydrogen bond may be sufficient to render a double-well potential ($x_0 > 0.693149$), but that the potential barrier will anyway be a rather small fraction of the bond strength. In these cases, one can hence expect a fast proton transfer above and through the barrier.

It is interesting to observe also the change of the curvature of a specific minimum under the influence of the second well. One has

$$\frac{V''}{D} = 2, \qquad \frac{W''}{D} = 2(1 - 4\alpha_0^2), \qquad \nu' = \nu\sqrt{1 - 4\alpha_0^2}, \tag{IV.89}$$

which shows that the minimum becomes flatter under the formation of a double well; i.e., there ought to be a red shift in the stretching frequency. This is a phenomenon which is frequently studied in the experimental literature; see, e.g., the excellent survey in Pimentel and McClellan (1960).

In a similar way, one can now study the attraction of an "outside" proton to two neutral molecules of different type. For the sake of simplicity, we will consider the superposition of two single wells having the same constant d but different depths D_1 and D_3. The asymmetry may then

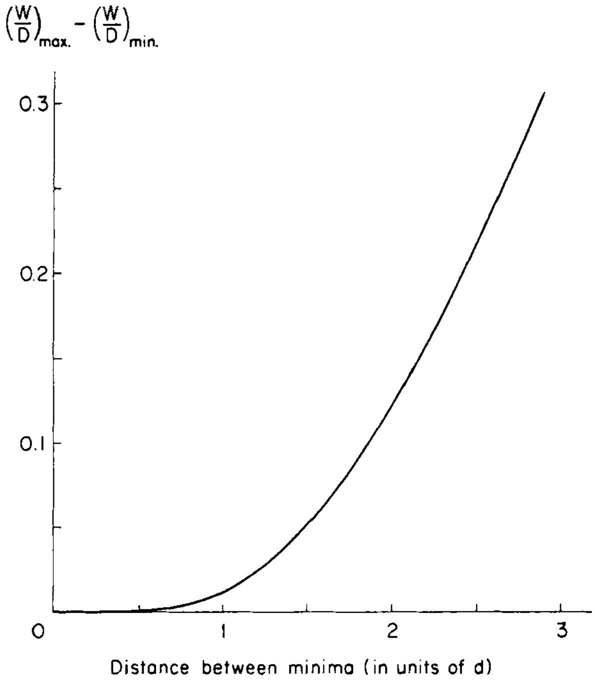


Fig. 48. The potential barrier as a function of the distance between the minima in a symmetric double-well potential obtained by superposition of two Morse potentials.

be characterized by the quotient $q = D_3/D_1$, and the potential takes the form:

$$W/D_1 = \{e^{-2x} - 2\,e^{-x}\} + q\{e^{-2(x_0-x)} - 2\,e^{-(x_0-x)}\}, \qquad \text{(IV.90)}$$

which is illustrated in Fig. 49. For $q = 0.1, 0.2$, and 0.3 the asymmetry is so high that there is only one single deep minimum for the values of x_0 investigated, whereas for $q = 0.4$ there starts developing a shallow minimum for $x_0 = 2.8$, which becomes deeper with increasing x_0.

Let us now consider an "inside" proton which originally belongs to one of the molecules under consideration. If this proton is removed from its equilibrium position a negative ion is formed, and the proton will be attracted by a Born–Mayer potential:

$$V(R) = -\frac{e^2}{R} + \frac{e^2}{R_0^2}\,de^{-(R-R_0)/d}, \qquad \text{(IV.91)}$$

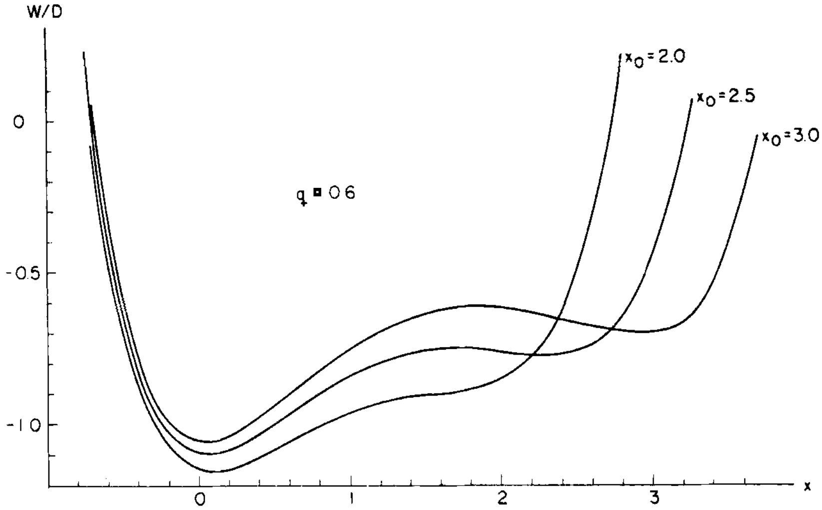


Fig. 49. Asymmetric double-well potentials obtained by superposition of two Morse potentials with different weight factors.

where $V(R_0) = -(e^2/R_0)(1 - d/R_0) = -D$ and e is the fundamental charge. We note that the Born–Mayer potential contains only two adjustable parameters, but that a third parameter may be introduced by treating the screening of the charge. As an alternative, one may also study the superposition of $-e^2/R$ and a Morse potential.

If the proton is now simultaneously attracted by another neutral molecule, the resulting potential may be represented by the superposition of a Born–Mayer potential and a Morse potential which is usually a highly asymmetric double well. If the second molecule B succeeds in stealing a proton from the first A, the ions B^+ and A^- are formed and one has the reaction

$$A + B \leftrightarrows A^- + B^+, \qquad \text{(IV. 92)}$$

e.g., $HCl + H_2O \rightleftarrows Cl^- + H_3O^+$. Experiences from dissociation theory tell us that such proton transfer may be both fast and rather complete. If, on the other hand, one considers only molecules of the same type, the probability for a proton transfer leading to the "self-dissociation"

$$A + A \leftrightarrows A^- + A^+ \qquad \text{(IV. 93)}$$

is usually very small. A typical example is given by water, which has the equilibrium constant $K = 10^{-14}$.

The hydrogen bonds of essential interest in biochemistry are those formed between nitrogen and oxygen, i.e., the bonds N : H ... : N, N : H ... : O, and O : H ... : O. In discussing these bonds, one can often to a certain extent utilize molecular data from inorganic chemistry. According to Pauling, for example, the average energy of the OH bond in water is 110.6 kcal/mole or 4.794 eV, the bond distance is 0.96 Å, and the stretching frequency is 3657.05 cm^{-1} with $\nu = 1.0963 \times 10^{14}$ sec^{-1}. The average energy of the NH bond in ammonia is 93.4 kcal/mole or 4.049 eV, the bond distance is 1.01 Å, and the stretching frequency is 3336.7 cm^{-1} with $\nu = 1.0094 \times 10^{14}$ sec^{-1}. It should be observed, however, that in biochemistry these bonds are greatly modified by the fact that the oxygen and nitrogen atoms involved usually belong to conjugated systems with mobile electrons, which are easily polarized by the protons. The form of the double-well potentials is hence dependent not only on the attraction between the protons and the electron lone pairs but also on the interaction between the protons and the π electrons. These effects will be discussed in greater detail in the following section.

In conclusion, we will study a general property of systems of hydrogen bonds and choose ice or water as an example. The crystal consists of discrete H_2O molecules in a tetrahedral arrangement. Each oxygen is surrounded by four other oxygens, and there is one proton along each O—O axis situated closer to one or the other of the two atoms. The proton arrangement is illustrated schematically in Fig. 50, and a similar linkage is characteristic also for water. The hydrogen bonds in such a structure are highly asymmetric, since each proton belongs to a neutral molecule, and the double-well potential acting on a proton is hence a superposition of a Born–Mayer potential from the OH^- ion and a Morse potential from the H_2O molecule. In such a hydrogen bond, the Boltzmann factor for the first tunneling level is very small, and the same is true for the total tunneling probability. However, if for some reason a proton tunnels through the

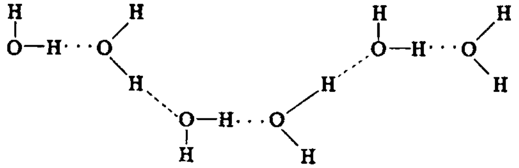


Fig. 50. Schematic illustration of the proton arrangement in the hydrogen bonds in ice.

barrier, an oxonium ion H_3O^+ is formed, and there will suddenly be two protons situated in symmetric O—O bonds. These protons have short tunneling times and will in turn influence their neighbors, etc. An addition of a proton to the system will hence start a kind of chain reaction, and this collective behavior is characteristic for systems having multiple hydrogen bonds. It will also be of essential importance in our discussion of the proton code of DNA.

V. Quantum Theory of DNA

A. Spontaneous Proton Tunneling in DNA

Let us consider a DNA molecule in its stable form during a period when it is biologically inactive, so that the genetic message is not undergoing replication or transcription. For this form, one can condense the essential features of the base pairing through hydrogen bonds, as shown in Fig. 51.

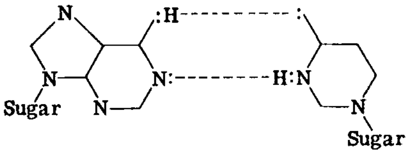


Fig. 51. Typical base pairing through hydrogen bonds in DNA.

There are at least two hydrogen bonds involved, and the problem of the genetic code is hence concerned with the question of the *motion and stability of two protons in a four-well potential* (see Fig. 52), i.e., one has to

Fig. 52. Asymmetric double-well potentials occurring in the hydrogen bonds of a base pair in DNA: (a) upper hydrogen bond, (b) lower hydrogen bond.

treat a quantum-mechanical two-body problem. Since the movements of the protons will further polarize the electron clouds and hence also change the potentials, a complete solution will undoubtedly be rather complicated.

For the sake of simplicity, we will first consider each one of the two double-well potentials separately. Since it is essential for the entire Watson–Crick model that the protons remain in their "normal" positions in the base pairs in order to represent a pure genetic message, it is necessary that the double-well potentials are asymmetric. Due to the ionic effect, the hydrogen bonds are certainly asymmetric, but the question is whether this asymmetry is large enough to give the protons the enormous stability which is characteristic for the genetic code.

However, even if the double-well potential in each hydrogen bond is highly asymmetric, one can expect that there will always be a small probability for proton transfer through and above the potential barrier in the middle of the bond in accordance with the results of the previous section (Löwdin, 1963, 1964a, c). Due to the fact that each base pair contains two or three hydrogen bonds, this leads to a collective phenomenon of the same type as discussed previously, e.g., in ice.

Since the bases are originally equally charged, the tunneling of one proton in one direction will very likely induce a tunneling of the other proton in the reverse direction to keep balance between the gross electric charges, so that

$$\begin{array}{ccc} :\mathrm{H}\overset{\rightarrow}{--}-----: & & :---------\mathrm{H}: \\ & \longrightarrow & \\ :-----\overset{\leftarrow}{--}\mathrm{H}: & & :\mathrm{H}--------: \end{array}$$

This "simultaneous" proton tunneling implies the base transitions:

$$\mathrm{A}\text{—}\mathrm{T} \longrightarrow \mathrm{A}^*\text{—}\mathrm{T}^*, \qquad \mathrm{G}\text{—}\mathrm{C} \longrightarrow \mathrm{G}^*\text{—}\mathrm{C}^*,$$

which lead to the production of *pairs* of tautomeric bases. If the hydrogen bonds get released in this position, the tautomeric forms will lead to errors in the next replication, i.e., to mutations.

One may consider these transitions as tunneling processes involving two protons shared between two equally charged partners and, since there are no "ionic" effects, one could anticipate that the double-well potential in the two-proton configuration space would have a rather symmetric character. However, mutations are exceedingly rare events, and even this double-well potential must hence be sufficiently asymmetric to give the genetic code the necessary stability. In this connection, one could perhaps understand why nature has chosen each base pair to consist of different partners—one purine and one pyrimidine base. In the following, we will investigate how this asymmetry is achieved. For this purpose, it is necessary to know the electronic structure of the base pairs and the DNA molecule in greater detail.

B. Electronic Structure of DNA

Since even the finer details of the electrostatic potentials acting on the two protons may be of importance in connection with the tunneling effect, it is highly desirable to determine the electronic structure of DNA with a great deal of accuracy. Since the knowledge of this electronic structure is of essential importance also in many other connections, a great deal of effort is being devoted to improving the methods available and the computational technique.

In classical chemistry, a conjugated system is an essentially planar molecule characterized by an alternating series of single and double bonds. In the modern description developed by Hückel (1931a, b, 1932), the electronic structure consists of a σ skeleton associated with the single bonds and a π electron cloud corresponding to the conventional double bonds. So far most of the attention has been devoted to the mobile π electrons and their chemical and physical importance.

The "Hückel scheme" is a special form of the molecular orbital method applied to aromatic compounds and conjugated systems in general. It is assumed that each π electron moves in the average field coming from the nuclei, the σ skeleton, and all the other π electrons and obeys a one-electron Schrödinger equation of the form:

$$H_{\text{eff}}\Psi_k = \varepsilon_k \Psi_k \tag{V.1}$$

where H_{eff} is the so-called effective Hamiltonian. By introducing the atomic p_z orbitals as a basis $\{\phi_\mu\}$ and assuming the existence of the expansion $\Psi_k = \sum_\nu \phi_\nu c_{\nu k}$, one obtains $\langle \phi_\mu | H_{\text{eff}}\ \Psi_k - \varepsilon_k \Psi_k \rangle = \sum_\nu \langle \phi_\mu | H_{\text{eff}} - \varepsilon_k | \phi_\nu \rangle c_{\nu k} = 0$. Introducing the matrix elements of the Hamiltonian, $H_{\mu\nu} = \langle \phi_\mu | H_{\text{eff}} | \phi_\nu \rangle$ with the special notations

$$\alpha_\mu = \langle \Phi_\mu | H_{\text{eff}} | \Phi_\mu \rangle, \qquad \beta_{\mu\nu} = \langle \Phi_\mu | H_{\text{eff}} | \Phi_\nu \rangle, \tag{V.2}$$

and the metric matrix $\Delta_{\mu\nu} = \langle \phi_\mu | \phi_\nu \rangle$, one obtains the system of linear equations

$$\sum_\nu (H_{\mu\nu} - \varepsilon \Delta_{\mu\nu}) c_{\nu k} = 0, \tag{V.3}$$

with the secular equation $\det\{H_{\mu\nu} - \varepsilon \Delta_{\mu\nu}\} = 0$ determining the orbital energies ε_k. After determining the coefficients $c_{\nu k}$, one can calculate the charge order q_μ of atom μ and the bond order $p_{\mu\nu}$ of the bond μ—ν by the relations:

$$q_\mu = \sum_k c_{\mu k} c_{k\mu}^+, \qquad p_{\mu\nu} = \tfrac{1}{2} \sum_k (c_{\mu k} c_{k\nu}^+ + c_{\nu k} c_{k\mu}^+), \tag{V.4}$$

where k is summed over all occupied orbitals Ψ_k, taking doubly occupied orbitals twice. The importance of the charge and bond orders as molecular indices has been emphasized by many authors (Hückel, 1931a, b, 1932; Wheland and Pauling, 1939; Wheland, 1942; Longuet-Higgins and Coulson, 1947; Coulson and Longuet-Higgins, 1947a, b, 1948). In the standard Hückel approximation, one considers only interaction between nearest neighbors, and the matrix elements α and β are treated as semi-empirical parameters; in heterocyclics, the differences in the α_μ values are further related to the electronegativity differences between the atoms involved (Daudel *et al.*, 1959a, b).

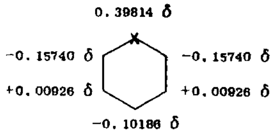
The parameter values are often expressed in the form $\alpha_\mu = \alpha + \delta_\mu \cdot \beta$, $\beta_{\mu\nu} = \eta_{\mu\nu} \cdot \beta$, where α and β are the corresponding quantities in the benzene molecule which are used as reference quantities.

In their encyclopedic study of the conjugated systems in biochemistry, the Pullmans have consistently used the Hückel approximation and for a survey of their results we will refer to their recently published book (B. and A. Pullman, 1963). Experiences from the quantum-mechanical study of many types of conjugated systems indicate, however, that in order to get a full and reliable understanding of the electronic structure of such systems in their ground state and excited states it is perhaps necessary to go beyond the Hückel approximation. We will return to this question after a discussion of the stability of the genetic code in DNA based on Hückel data.

1. Charge Orders of Purine and Pyrimidine Bases

The Hückel scheme is particularly useful in studying heterocyclics, and it was early found that the charge and bond orders could be found directly without solving the secular equation by means of considerations related to perturbation theory. This will be illustrated here by some simple examples which may facilitate the understanding of the charge structure of the purine and pyrimidine bases.

a. Heteroatoms in the Ring; Hückel Parameters. A benzene molecule has six π electrons, and the charge orders are all equal to 1. If a heteroatom (X) characterized by the parameter $(\alpha + \delta\beta)$ is introduced into the ring, the charge orders will undergo certain perturbations proportional to δ which for the self-, *ortho*, *meta*, and *para* positions have the coefficients 43/108, −17/108, 1/108, and −11/108, respectively (Wheland and Pauling, 1939; Löwdin, 1953). This may be illustrated by:



(V.5)

By using perturbation theory (Löwdin, 1953), one can show that these changes in first order are *additive* so that, if a molecule contains several heteroatoms, one can simply add the charge order perturbations associated with the isolated heteroatoms.

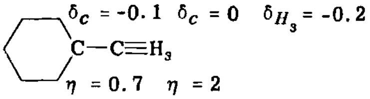
Following Wheland and Pauling (1939), one has generally assumed that, if a heteroatom is associated with an α parameter δ in its own position, it induces also a small perturbation of the order 0.1δ in the α parameters of the neighboring positions. By superimposing the three perturbation patterns resulting from this hypothesis, one obtains changes in the charge orders which, for the various positions in order, are characterized by the coefficients 11/30, −7/60, −1/60, −1/10, respectively. The main importance of this inductive effect is the change in sign of the coefficient for the metaposition from 1/108 to −1/60.

Using the bond length of benzene ($a = 1.40 \times 10^{-8}$ cm) and the value of the fundamental charge ($e = 4.803 \times 10^{-10}$ esu), one finds by vector addition that a perturbation δ gives rise to a contribution to the dipole moment from the π electrons corresponding to 2.24 δ Debye units. If the inductive effects are taken into account, this figure is increased to 2.46 δ Debye units. These data have been used to determine certain δ values from empirically known dipole moments. Since the δ values are further related to the electronegativities, this approach can be used to predict dipole moments for molecules which have not yet been experimentally investigated.

The first set of δ values was determined by Wheland and Pauling (1939) from dipole moments, and a characteristic value is $\delta_N = 2$ for a ring nitrogen. This value was later reduced to $\delta_N = 0.6$ (Löwdin, 1953b) and the entire scale changed proportionally (Orgel *et al.*, 1951). In their work on heterocyclics of importance in biochemistry, the Pullmans have chosen a still lower value, $\delta_N = 0.4$, but they have at the same time neglected the inductive effect (except for some very high δ values). In order for us to be able to compare our results with the Pullmans' data, we will use their parameter values which are specified in Eqs. (V.6). (Note that, for the methyl group, a misprint has been corrected.)

$$\begin{array}{lll} \text{ring nitrogen:} & \delta_N = 0.4 & \eta_{CN} = 1 \\ & \delta_{NH} = 1 & \eta_{CN} = 0.9 \\ \text{amino nitrogen:} & \delta_{NH} = 0.4 & \eta_{CN} = 1 \\ & \delta_{NH_2} = 1 & \eta_{CN} = 0.9 \\ \text{keto oxygen:} & \delta_O = 1.2 & \eta_{CO} = 2 \\ & \delta_{OH} = 2 & \eta_{CO} = 0.9 \end{array} \qquad \text{(V. 6)}$$

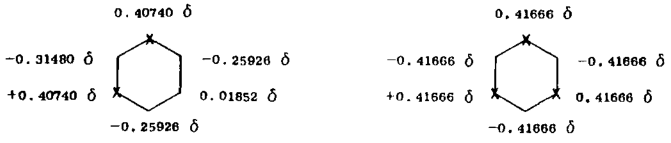
methyl group:



The methyl group contributes two π electrons through hyperconjugation.

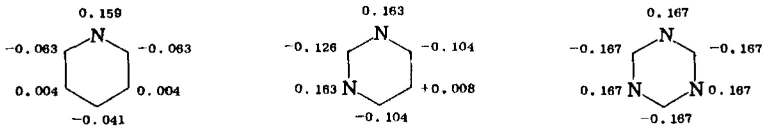
The parameter values are chosen to give good agreement between the theoretical results of the Hückel scheme and a large number of *empirical data.*

As an example of the additivity rule, we will now consider a benzene ring with two or three heteroatoms of the same type. By superimposing the perturbation patterns, one obtains for instance:



(V.7)

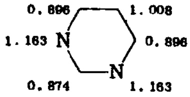
Putting $\delta_N = 0.4$, one gets the following charge order perturbations for pyridine, pyrimidine, and pyrazine:



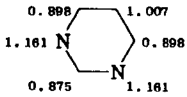
Pyridine Pyrimidine *s*-Triazine

(V.8)

These figures are in good agreement (±0.002) with the data obtained by solving the secular equation. For pyrimidine, one has consequently the following total charge orders:



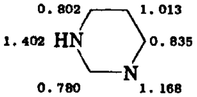
Charge orders derived
by additivity rule



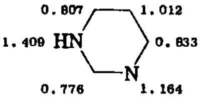
Charge orders derived
by solving secular equation

(V.9)

In order to investigate the influence of a proton on the π electron cloud, we will now substitute N:H instead of N in pyrimidine and use the Hückle parameters $\delta_{\mathrm{NH}} = 1$ and $\eta_{\mathrm{CN}} = 0.9$ in (V.6). In a first approximation, the charge orders are independent of small changes in η and, putting $\delta = 1$ and $\delta = 0.4$ into (V.5) and adding the perturbation patterns, one obtains:



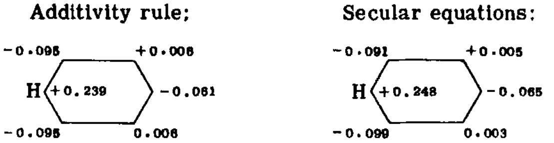
Charge orders derived
by additivity rule



Charge orders derived
by solving secular equation

(V.10)

It is remarkable that, even for such a large δ value as $\delta = 1$, the additivity rule gives a result in fair agreement with the charge orders obtained by solving the secular equation. A comparison between (V.9) and (V.10) shows further that a proton has a perturbation pattern of the following type:


Additivity rule;

Secular equations:

(V.11)

Polarizing effect on π electron cloud
of proton approaching a ring nitrogen

The polarizing effect of a proton approaching a ring nitrogen is hence appreciable and plays an important role in the discussion of the hydrogen bonds. A study of the orbital energies indicates further that the total π electron energy decreases from 4.433 β to 4.691 β when the proton gets attached.

b. Effect of a Substituent. Next we will consider the changes of the charge orders in a benzene ring caused by a *substituent* which adds another atom (X) and two π electrons to the conjugated system. If a perturbation δ_X is put into this position (see Fig. 53), one obtains the charge orders given in Table V. The data in these tables are sufficient to permit interpolation of the charge orders correct to three decimals. Using perturbation theory,

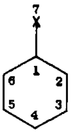


Fig. 53. Benzene ring with a substituent.

TABLE V

CHARGE ORDERS OF A SUBSTITUTED BENZENE MOLECULE WITH EIGHT π ELECTRONS

$\eta_{1,7} = 0.9$

Atom	$\delta = 0$	$\delta = 0.5$	NH$_2$ $\delta = 1.0$	$\delta = 1.5$	OH $\delta = 2.0$
7	1.622 084	1.770 485	1.850 582	1.896 418	1.924 470
1	1.000 000	0.960 760	0.946 960	0.943 499	0.944 155
2	1.125 972	1.095 209	1.074 288	1.060 000	1.049 911
3	1.000 000	0.997 506	0.997 082	0.997 223	0.997 489
4	1.125 972	1.083 324	1.059 716	1.045 637	1.036 575

$\eta_{1,7} = 1$

Atom	$\delta = 0$	NH $\delta = 0.4$	$\delta = 0.5$	$\delta = 1.0$	$\delta = 1.5$	$\delta = 2.0$
7	1.571 429	1.704 575	1.729 522	1.820 101	1.873 792	1.907 338
1	1.000 000	0.961 423	0.955 361	0.938 027	0.932 902	0.932 997
2	1.142 857	1.116 957	1.111 388	1.088 544	1.072 308	1.060 561
3	1.000 000	0.997 491	0.997 188	0.996 603	0.996 706	0.996 986
4	1.142 857	1.105 106	1.097 963	1.071 577	1.055 278	1.044 569

$\eta_{1,7} = 2$

Atom	$\delta = 0$	$\delta = 0.5$	$\delta = 1.0$	O $\delta = 1.2$	$\delta = 1.5$	$\delta = 2.0$
7	1.250 000	1.391 409	1.510 937	1.552 548	1.607 556	1.683 607
1	1.000 000	0.931 482	0.881 923	0.866 352	0.849 313	0.829 925
2	1.250 000	1.234 368	1.216 384	1.210 696	1.197 913	1.180 250
3	1.000 000	0.996 241	0.994 047	0.994 722	0.992 954	0.992 555
4	1.250 000	1.215 889	1.186 277	1.174 318	1.161 394	1.140 856

one can give the corresponding analytic formulas but for the applications it is more convenient to use the tables (see Table V). Note that the numbering of the atoms is special for this section and should not be confused with the conventional numbering used in describing certain molecules (cf. Section II, D).

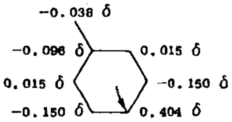
Of particular interest are the mutabilities $\pi_{\mu,\nu} = \partial q_\mu / \partial \delta_\nu$ (Coulson and Longuet-Higgins, 1947a, b, 1948a, b) which give the change of a charge order q_μ with respect to a change of a parameter δ_ν. Using the relation $q_\mu = \partial E / \partial \alpha_\mu$, one obtains the following symmetry theorem:

$$\frac{\partial q_\mu}{\partial \alpha_\nu} = \frac{\partial^2 E}{\partial \alpha_\nu \partial \alpha_\mu} = \frac{\partial q_\nu}{\partial \alpha_\mu} \tag{V.12}$$

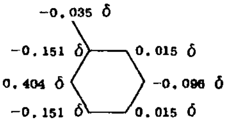
showing that $\pi_{\mu,\nu} = \pi_{\nu,\mu}$. From Table V, one can now derive numerical values for the mutabilities $\pi_{7,k} = \pi_{k,7}$. For $\eta_{7,1} = 0.9$ and $\delta = 1$, one obtains approximately:

$$\begin{gathered} \pi_{7,1} = -0.014, \qquad \pi_{7,2} = -0.035, \qquad \pi_{7,3} = 0, \\ \pi_{7,4} = -0.038, \qquad \pi_{7,7} = +0.122, \end{gathered} \tag{V.13}$$

which figures may be easily improved. By means of these quantities, it is now possible to calculate the change of the charge order in the 7th position by introducing a heteroatom in the ring. For example, if the perturbation δ is introduced in position 4, a charge order -0.038δ will be induced in position 7 according to (V.13), and there are then $(6 + 0.038\delta)$ mobile electrons in the ring which are distributed approximately in proportion to the charge orders of the ring alone. This gives a perturbation pattern of the type:



Perturbation δ in position 4

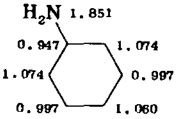


Perturbation δ in position 2

(V.14)

Similarly, one can construct the perturbation pattern for position 2, etc. Still better data may be obtained by studying the perturbation problem for each position separately.

The charge orders for aniline are given in Table V and can be illustrated by:



(V.15)

By superimposing the perturbation pattern of substituted pyrimidine according to (V.14) with $\delta = 0.4$, one obtains the charge orders for aminopyrimidine:

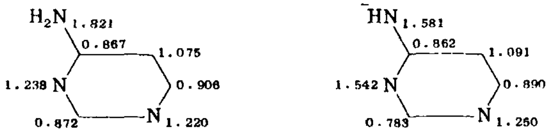
Additivity rule

Secular equation

(V.16)

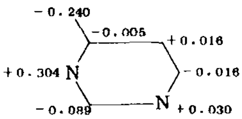
The agreement with the charge orders obtained from the secular equation is at least fair, and the use of the simple additivity rule can easily be improved.

Let us now consider the polarizing effect of a proton which approaches or leaves the substituent. For this purpose, a comparison of the charge orders of the two molecules is given:



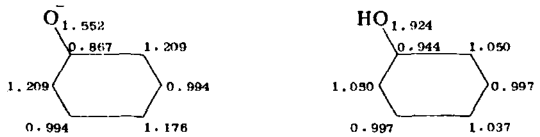
(V.17)

The removal of a proton from the amino group is hence associated with the following perturbation pattern:



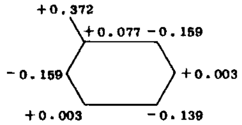
(V.18)

Comparison between the charge orders of the two molecules shown in



(V.19)

(V.19) gives the following perturbation pattern:



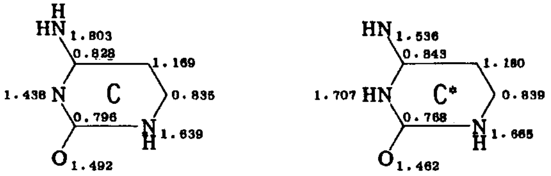
(V.20)

for the approach of a proton to a keto oxygen. In all cases, the polarizing effect of a proton on the π electrons is quite appreciable.

Even if the additivity rule is only approximate, it may be of practical importance in such applications where it is essential to construct molecules having a certain pattern of charge orders, e.g., in connection with "quantum pharmacology." Further work on this problem is in progress.

c. Pyrimidine and Purine Bases. By using the additivity principle, one can in this way derive the charge distributions for the pyrimidine bases occurring in DNA and their tautomeric forms. Except for small deviations associated with the use of perturbation theory, the results shown in (V.21) and (V.22) are identical with those which may be obtained by solving the secular equations:

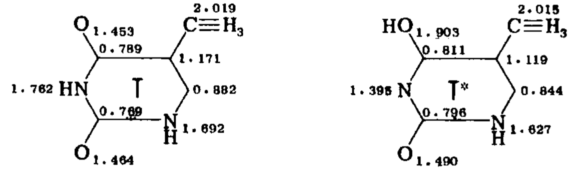
Cytosine



$\epsilon_\pi = 10\alpha + 16.096\,\beta$ $\epsilon_\pi = 10\alpha + 15.962\,\beta$

(V.21)

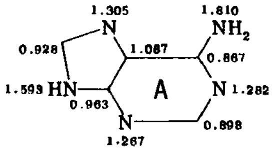
Thymine



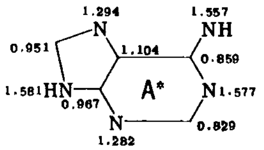
$\epsilon_\pi = 12\alpha + 22.512\,\beta$ $\epsilon_\pi = 12\alpha + 21.798\,\beta$

(V.22)

The purine bases may be treated similarly. Starting out from the charge orders for indole and benzimidazole and adding the perturbation patterns for the pyrimidine bases, one can get a fair idea of the charge distributions also in the purine bases. The results shown in (V.23) and (V.24) have been obtained by solving the secular equations:

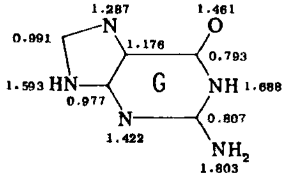


$\epsilon_\pi = 12\alpha + 17.214\beta$

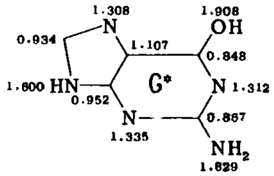


$\epsilon_\pi = 12\alpha + 16.940\beta$

(V.23)



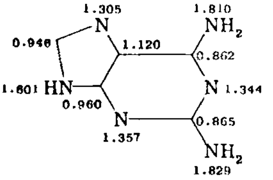
$\epsilon_\pi = 14\alpha + 22.094\beta$



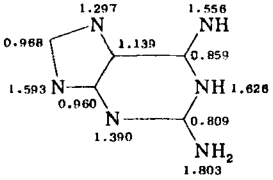
$\epsilon_\pi = 14\alpha + 21.482\beta$

(V.24)

The results given in (V.21–V.24) are essentially the same as those previously reported by the Pullmans (A. and B. Pullman, 1961a, 1963), but we have here concentrated our interest on the normal forms and those tautomeric forms as may occur in connection with errors in the genetic code. We note that the π electron energies of all these tautomeric forms are higher than the energies for the normal forms and that the differences are -0.134β,

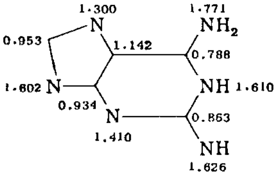


$\epsilon_\pi = 14\alpha + 19.614\beta$



$\epsilon_\pi = 14\alpha + 19.392\beta$

(V.25)



$\epsilon_\pi = 14\alpha + 19.356\beta$

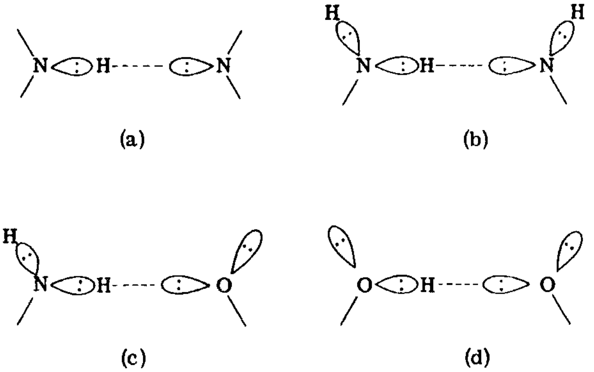
-0.714β, -0.274β, and -0.618β for C, T, A, and G, respectively. For benzene, one has $\beta = -18$ kcal/mole $= -0.78$ eV, and for a more complete discussion of the choice of proper β values we recommend the Pullmans' book (B. and A. Pullman, 1963).

In conclusion, a few words should also be said about an important adenine analog, namely 2, 6-diaminopurine. This molecule is of importance, since it could form *three* hydrogen bonds with thymine. For the charge orders of the normal form and the two tautomeric forms of particular interest, one finds the result shown in (V.25).

Even in this case, the normal form has the lowest π electron energy. One may wonder why nature has chosen adenine instead of 2, 6-diaminopurine as a building stone of DNA, and we will later return to this problem.

2. *Stability of Hydrogen Bonds in DNA*

Some of the most fundamental hydrogen bonds in biochemistry are the four types shown in Fig. 54. Topologically one can go from (b) to (c) and then to (d) by successively dropping the protons participating in the covalent bonds into the nitrogen nuclei, changing them to oxygen nuclei. Considering only the nuclei and the σ skeleton, the three forms (a), (b), and (d) look rather symmetric, but one should remember that, if the total system is neutral, the proton cannot be removed from the left-hand molecule without giving rise to a negative ion, and this "ionic effect" renders a great deal of asymmetry to the hydrogen bond involved. On the other hand, if the right-hand molecule obtains a single negative charge, the double-well potential becomes more symmetrical and the probability for proton transfer increases appreciably.



(a) (b)

(c) (d)

Fig. 54. The fundamental hydrogen bonds occurring in biochemistry.

The hydrogen bonds between the base pairs in DNA are of types (a) and (c), and we will now study their stability particularly with respect to proton transfer. In the discussion, it is convenient to distinguish between two cases which approximately correspond to spontaneous and induced proton transfer. In the former case the bases in a pair are in their normal and electrically neutral form, whereas in the latter case the bases are unequally charged. In this section we will discuss the normal case.

In determining the properties of the hydrogen bonds, the mobile electrons are going to play a fundamental role. In a preliminary study (Löwdin, 1963) of the normal forms of the base pairs, the author pointed out that the differences in charge orders across the bonds were by far the smallest in the middle bond of the G—C pair and that proton transfer should hence most likely occur in this bond (see Fig. 55). During the last year, several attempts to calculate actual shape of the double-well potentials associated with the hydrogen bonds in DNA have been made by the Uppsala group, and this work is still in progress.

The potential felt by a proton in the middle hydrogen bond of a G—C pair has been evaluated by Ladik (1963b) under the assumption that both σ and π electrons stay in their original state; i.e., the polarizing effect of the proton is neglected in a first approximation. This study is based on π electron data taken from a semiempirical MO—LCAO—SCF calculation of the G—C pair (Rein and Ladik, 1964). The charge orders are here still more "asymmetric" across the hydrogen bonds than the values obtained by the Hückel method reported above. Ladik's result is given in Fig. 56.

In evaluating the potential acting on the proton, two simple models have been used: in model 1 the two electron lone pairs are assumed to be localized in sp^2 hybrids of the nitrogen atoms directed opposite each other along the hydrogen bonds, whereas in model 2 the four σ electrons are assumed to fill the two lowest molecular orbitals of a simple type associated with the NH ... N system. In both models, the potential curves are highly asymmetric but otherwise are rather different, and it seems as if model 1 would give the most reasonable values. The difference between the minima is 2.72 eV, whereas the tunneling barrier is $V_0 = 0.68$ eV. The results are apparently highly sensitive to the assumptions about the electron distribution. The asymmetry depends on the previously discussed "ionic" effect, and it is evident that such a hydrogen bond would have a great deal of stability.

However, the polarizing effect of the proton is going to change this picture partly and, in the next refinement, it is necessary to study the

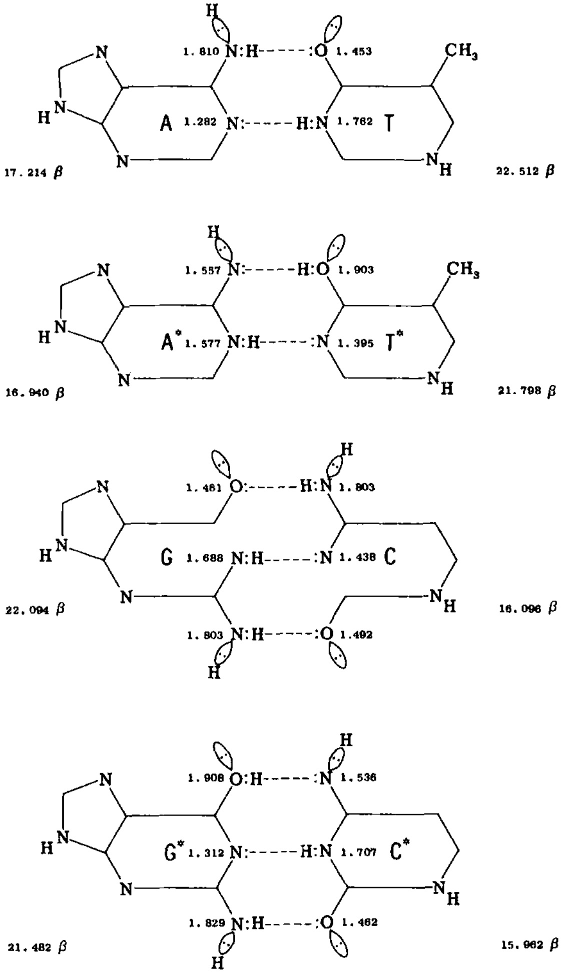


Fig. 55. Charge orders of the π-electrons for the base pairs in DNA in their normal and tautomeric forms. Note the strong polarizing effect associated with the motion of a proton along a hydrogen bond. The asymmetry of the charges on the opposite sides of a hydrogen bond will be slightly diminished if the "conjugating power" of the bond is taken into account.

influence of the proton and its movement on the σ and π electrons. A first attempt to study this problem in Uppsala has been made by Rein and Harris (1964). In investigating the middle hydrogen bond N:H ... :N of the G—C pair, they have determined the distribution of the twenty-four π electrons and the four σ electrons of the hydrogen bond region for a specific protonic position by means of the MO—LCAO—SCF method in a

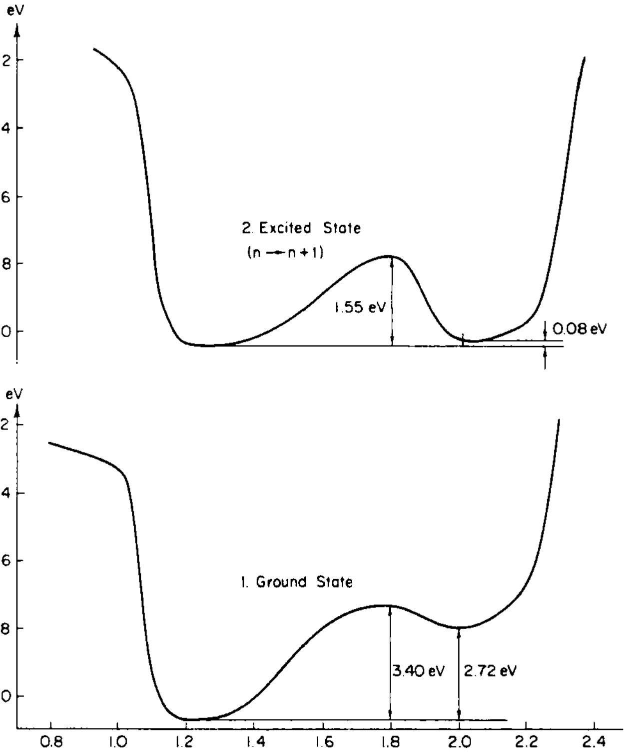


Fig. 56. Double-well potentials of the middle hydrogen bond N: H ... H of the G—C base pair in its ground state and first excited state according to Ladik (1963). In his model, the four σ-electrons are assumed to form two sp^2 lone pairs on each nitrogen atom involved. The mobile electrons are assumed to stay unpolarized.

somewhat simplified form. By repeating the calculation for twelve subsequent protonic positions along the bond, they have obtained an effective protonic potential which is given in Fig. 57. The asymmetry between the minima is here 1.4 eV, whereas the tunneling barrier is $V_0 = 0.4$ eV. Even if the asymmetry is essentially diminished by the polarization effects, it is still large enough to give the bond sufficient stability. According to Table IV, the Boltzmann factor for the lowest tunneling level and $T = 310°\mathrm{K}$ is $B = 10^{-23}$. The double-well potential obtained is illustrated in Fig. 57.

In the previous discussion, we have emphasized the importance of simultaneous (or almost simultaneous) proton tunneling in opposite directions in two neighboring hydrogen bonds, since such a process avoids the formation of ions and may hence lead to a final state which is energetically not too high in comparison with the original state. One may get a first rough idea of this phenomenon by considering the charge orders for the base pairs of DNA obtained by using the Hückel data discussed in the previous section (see Fig. 55). A comparison between the normal and tautomeric forms shows that the polarizing effect of the motion of a proton

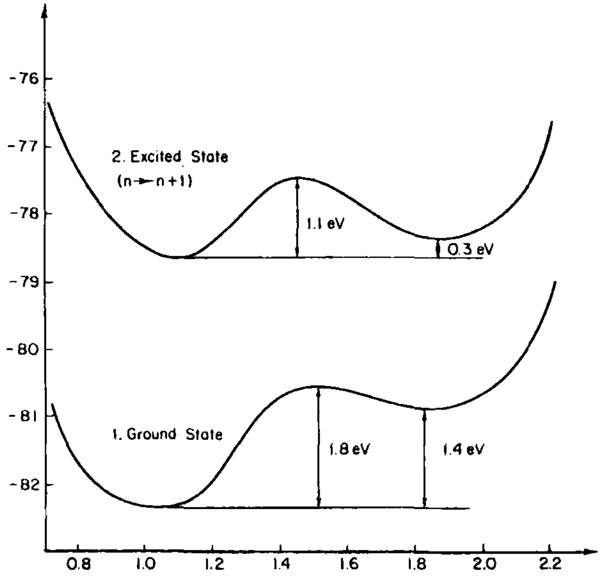


Fig. 57. Double-well potential of the N—H ... N hydrogen bond of the G—C base pair according to Rein and Harris (1964). The potential has been calculated by an LCAO—MO—SCF approximation considering explicitly the twenty-four π-electrons and four σ-electrons of the hydrogen bond region. The delocalization effect was taken care of by solving the SCF equations repeatedly for twelve subsequent protonic positions in the hydrogen bond region.

along a hydrogen bond is very large—it is actually so large that one wonders whether the special Hückel parameters in (V.6) may overemphasize this effect. The asymmetry of the charge orders on opposite sides of a hydrogen bond is slightly diminished if one introduces the "conjugating power" of the hydrogen bond, e.g., by following the simple recipe given by the Pullmans (B. and A. Pullman, 1963), but the charge order transported across the bond is still only of the order of magnitude ±0.050. In the transition from a normal to a tautomeric base pair, the total π electron energy increases and, for the A—T and G—C pairs, one obtains the following figures:

$$\begin{aligned} -0.714\,\beta - 0.274\,\beta &= -0.988\,\beta \quad \text{(AT)},\\ -0.618\,\beta - 0.134\,\beta &= -0.752\,\beta \quad \text{(GC)}, \end{aligned} \tag{V.26}$$

respectively, i.e., 0.771 eV and 0.586 eV. It may be worthwhile to observe that, for the "base pair" consisting of 2, 6-diaminopurine and thymine, the corresponding figure is only $-0.356\beta = 0.278$ eV, which may be one of the reasons why nature has not utilized this base pair as a building stone in DNA.

It should be observed in this connection that the base pairs A*—C and A—C* are transformed into each other by double proton tunneling, and that the same applies also to the pairs G*—T and G—T*. Total π electron energies are given below:

A*—C	A—C*	G*—T	G—T*
$22\alpha + 33.036\beta$	$22\alpha + 33.176\beta$	$26\alpha + 43.994\beta$	$26\alpha + 43.892\beta$

and favor the forms A—C* and G*—T, in which the partners A and T are normal. It would be of essential interest to calculate the differences in total energies, since the balance between the two forms is going to influence the outcome of tautomeric errors in the genetic code and may favor the stability of one of the normal base pairs, probably the A—T pair.

In order to get detailed information about double proton tunneling, it is necessary to evaluate the energy of the base pairs as a function of the position of the two protons when the polarizing effect of their motion on the σ and π electrons is taken into full account (see Fig. 58). Using the previously mentioned MO—LCAO—SCF method for the σ and π electrons. Rein and Harris are now evaluating the shape of the energy surface for the G—C and A—T pairs utilizing a network of points.

Considering the energy difference 1.4 eV in Fig. 7c for a single proton tunneling, it seems reasonable that the energy difference should be diminished to about 0.6–1 eV after the tunneling of the second proton and return to neutral bases; the value actually found is 0.7 eV. This implies that the Boltzmann factor for the lowest tunneling level and $T = 310$°K will lie in the region between 10^{-10} and 10^{-16}. Actual experience shows that the probability per base pair and generation for a spontaneous mutation is of the order of magnitude 10^{-8} to 10^{-11} (Freese, 1962). Considering the difficulty of carrying out the quantum-mechanical calculations with sufficient accuracy, such an agreement is certainly too good to be true, and for a rather long time, one will probably have to be satisfied if one obtains reasonable agreements as to relative mutation rates.

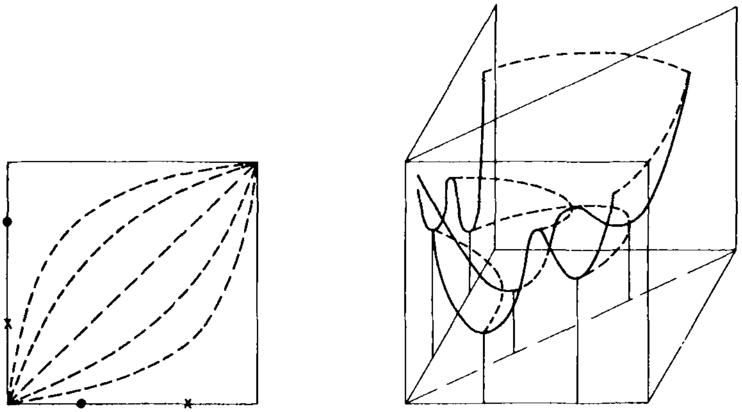

(a) (b)

Fig. 58. (a) Pathways for double proton tunneling. (b) Energy surface associated with double proton tunneling.

In connection with more detailed studies of the base pairs, one should be particularly careful about definite conclusions as to absolute spontaneous mutation rates as long as the calculations are not carried out with sufficient accuracy on an "*ab initio*" basis. The Hückel scheme, for instance, is based on the use of semiempirical parameters of the type δ, η, and β, which may be chosen in different ways—in the literature, for example, one has suggested for nitrogen the δ values $\delta_N = 2$, $\delta_N = 0.6$, $\delta_N = 0.4$, and even $\delta_N = 0.2$. Usually the properties of a molecule are not so sensitive to the choice of these Hückel parameters, which explains the success one has previously had with this approach in the theory of conjugate systems, but in the case of proton tunneling the situation is entirely different. A change of a charge order by 0.01 may actually influence the Boltzmann factor for the lowest tunneling level by a full factor 10, and by a small change of the δ scale one can hence achieve a large adjustment of the Boltzmann factor B for the lowest tunneling level and hence also for the spontaneous mutation rates. This is certainly not a proper method for fixing the δ scale, but the discussion shows the high degree of accuracy needed in the quantum-mechanical calculations to get reliable results as to proton tunneling.

In spite of these difficulties, certain of our preliminary results are at least qualitatively correct. It seems likely that hydrogen bonds of type (a) and (c) between purine and pyrimidine bases in neutral form show a sufficiently high degree of asymmetry due to the ionic effect to be stable with respect to single proton transfer. However, if the right-hand base gets a single negative charge (or the left-hand base a single positive charge), the double-well potential in (a) becomes much more symmetric and permits apparently an appreciable amount of single proton tunneling to the right-hand position. Our preliminary study indicates further that, with respect to double proton tunneling, the A—T pair is probably considerably more stable than the G—C pair. The final conclusion is that a good treatment of the hydrogen bonds in biochemistry is possible first after an essential improvement of the present quantum theory of the conjugated systems.

3. *Refinements of the Theory*

The Hückel scheme has recently been refined by extending the so-called self-consistent field ideas to exact form. This has been achieved by considering a formal solution to the many-electron Schrödinger equation based on a simple partitioning technique equivalent to infinite-order perturbation theory. The reaction operator obtained in this way is utilized to construct an exact self-consistent field scheme (Löwdin, 1962c) which is a direct generalization of the Hückel method. The scheme may be used for *ab initio* calculations by means of electronic computers but seems still more appropriate for semiempirical discussions and for the study of fundamental concepts. This approach has so far been used to investigate the problem of the change of bond orders and bond lengths under addition and substitution, and to study the connection between the Hückel parameters and Pauling's concepts of electronegativity.

Symmetry properties and particularly the total spin could further be conveniently treated by the projection operator technique (Löwdin, 1962f). In addition to the Hückel scheme, the method using "different orbitals for different spins" could also be utilized for treating the correlation problem, particularly in the alternant conjugated systems involved (Löwdin, 1954, 1955, 1959, 1960, 1962a; Paunez *et al.*, 1962). Full attention should be paid to the nonorthogonality problem (Löwdin, 1950, 1956).

Such an investigation of the detailed electronic structure of DNA should start by considering the single bases, then base pairs, then a series of base pairs, etc., (Ladik, 1960, 1961, 1963a) until one obtains a reasonable model of DNA (Hoffmann and Ladik, 1961, 1964a,b). One can utilize the fact that the electronic structure of the bases involved remains rather stable and that the actual changes may be calculated by means of the theory for "localized perturbations" (Löwdin, 1962c; Ali and Wood, 1962). In addition to the π electron system of the bases, a great deal of attention should be devoted to the σ electron framework, and in particular the character of the "lone pairs" should be studied in full detail. Work along these lines is now in progress in Uppsala.

It should be observed that the problem of the two protons in the four-well electrostatic potential is actually a quantum-mechanical two-particle problem, and some of the methods previously developed for treating such systems could perhaps hence be utilized also in this connection. In a first approximation, the problem could be treated according to the independent-particle model, but, since the protons repel each other strongly, it seems desirable to introduce a certain amount of correlation between their motions. The previous experience gained in treating two-particle systems will here certainly turn out to be useful.

The polarization of the electronic cloud in connection with the motion of the protons is finally another important aspect of the problem which ought to be given full consideration as well as the question of the influence of the thermal vibrations of the bases on the properties of the hydrogen bonds. Even in this more complicated case, one tries often to reduce the problem to the question of the motion of a point on a multidimensional potential energy surface, but the point is then no longer associated with a single proton but represents the state of the system in a configuration space. The formulation and solution of this problem require a great deal of effort in extending the present methods for the application of quantum theory to many-particle systems, but the problem is certainly one of the most challenging and stimulating ones in the entire field, and one should search also for new approaches.

C. Induced Proton Tunneling in DNA

The treatment of the problem of the proton transfer in the normal base pairs of DNA shows that all the hydrogen bonds are probably exceedingly stable, and that the transfer rates are so small that it is hard to say whether they may become recognizable. If this problem is hence highly delicate, the problem of the induced proton tunneling is of a different type and involves effects which are larger by many orders of magnitude.

Here we will discuss only such induced processes as are caused by the fact that the two bases within a pair have obtained unequal charges. Under these conditions, the proton transfer goes essentially in one direction, i.e., towards the base which is negatively charged in comparison to the other:

$$\begin{array}{ccc} :H\overset{\rightarrow}{\cdot\cdot}\cdots: & & :\cdots\cdots H: \\ & \longrightarrow & \\ :\cdots\cdots H: & & :\cdots\cdots H: \end{array}$$

This leads consequently to transitions of the type A—T → A^+—T^- or A^-—T^+, and the type G—C → G^+—C^- or G^-—C^+. In this case one obtains two ionic tautomeric forms which do not combine with the normal bases and which, if they remain in this form at the moment of replication, may even cause deletions in one strand of the base sequence leading to irreversible mutations.

It should be noted that, if an extra charge is added to one of the bases in a pair, the asymmetry of the hydrogen bonds may be completely changed (see Fig. 59); so that the proton transfer rates may go up by many orders of magnitude. In this connection one should consider not only the tunneling through but also the passage over the barrier, if there remains one.

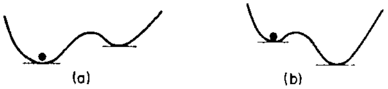


Fig. 59. Change of character of a double-well potential if an extra charge is added to (or removed from) one of the partners involved: (a) normal, (b) negative charge added on the right-hand base.

One may wonder under what conditions the bases within a pair get unequal charges. Through electronic donor-acceptor reactions with other molecules, an electron may be added to (or removed from) the π electron cloud of one of the bases. The reacting molecule may enter the "deep groove" of the DNA molecule or may be "sandwiched" between two successive base pairs. One electron is further often removed from the π part of one of the bases through the (direct or indirect) effect of ionizing radiation.

The effect of weak UV-radiation has recently been studied by Rein and Ladik (1964), who pointed out that, in a G—C pair, such radiation may cause a π-π^* transition in which an electron leaves the highest occupied orbital of guanine and enters the lowest unoccupied orbital of cytosine. This excitation process is hence connected with a charge transfer within the molecule which will influence the shape of the double-well potentials. The potential for the middle bond of the G—C pair has been evalauted (Ladik, 1963; Rein and Harris, 1964a), and the results are illustrated in Figs. 7b and 7c. It turns out that the asymmetry of the bond characteristic for the ground state is largely removed in the excited state, and Rein and Harris obtain an energy difference between the minima of 0.3 eV and a tunneling barrier $V_0 = 0.8$ eV. According to Tables II and IV, the tunneling time is approximately 10^{-5} sec and the Boltzmann factor for the lowest tunneling level is $B = 10^{-5}$. Provided that the lifetime of the excited state is long enough, the weak UV-radiation will hence give rise to a considerable mutation rate.

An important process is probably the addition of a proton to one of the purine bases through bonding to one of the extra electron lone pairs available at the N-3 and N-7 positions, which means that proper attention should be given also to proton reactions. It should be mentioned finally that the double-well potential may be disturbed through additional electrostatic potentials from outer sources, dipole double layers, etc. The dimensions may also be changed through external pressure, ultrasound waves, etc. Other interesting problems to be studied include the effect of radiation in resonance with the proton frequencies associated with the double-well potential as well as the effect of strong magnetic fields on the proton spins, and we will return to these questions later.

VI. Some Remarks on the Replication of DNA, the Formation of RNA, and the Coding Problem

It should be observed that, if the determination of the structure of the stationary form of DNA is a "static" problem, the question of the replication of DNA and the formation of RNA represents "dynamic" problems of a still higher degree of difficulty. In connection with the problems of "reading and transcribing the genetic code," it is hence much harder to reach conclusive results, and many of the arguments have to be rather speculative, based on the interpretation of the results of the current experiments in genetics and in cellular and molecular biology. In this section, we will try to give a review of some of the basic principles involved together with a brief summary of the mechanisms proposed by various authors. It should further be stressed that, even though the discussion of the biological functioning of DNA in the later sections certainly depends on the mechanisms for "reading the code," the main interpretations still depend essentially only on the general copying principle—i.e., the complementarity concept for the nucleotides based on the proton-electron pair code for hydrogen bonding—and not on the details of the mechanism. After these words of precaution, we can now proceed to the details.

A. Replication of DNA

In the original Watson–Crick model of DNA, it was simply suggested that, before the cell division, the two strands of the double helix would separate and that each one would then build its own complement leading to two new double helices identical with the original one. Even though the basic principle is exceedingly simple, the actual mechanism involved may be rather complicated. The following three physical laws valid in both classical physics and quantum theory have undoubtedly to be satisfied, namely the conservation laws for the total energy, momentum, and angular momentum.

The energy law is of essential importance. The main argument against the complete strand separation hypothesis in the Watson–Crick model comes from the fact that it would require a very large amount of energy to break all the hydrogen bonds in the DNA double helix simultaneously. It seems instead more feasible to assume a mechanism where the synthesis of the complementary strands goes parallel with the strand separation in the original helix, so that at least part of the energy gained in the formation of the four to six new hydrogen bonds could be used to break the two to three hydrogen bonds in another base pair in the old double helix. Since the net result is that one has formed two double helices from one double helix and free nucleotides, the energy totally released depends on the number of hydrogen bonds formed in one helix with a correction for the energy amount involved in linking the sugar-phosphate backbones together.

The law of conservation of momentum implies that the mass centrum of the total system (old double helix plus nucleotide building material, enzymes, etc.) will not be changed under the replication process unless there is some interaction with the boundaries. It is clear that, in a process where the synthesis of the two new strands goes parallel with the breaking up of the old helix, there will undoubtedly be a complicated winding-unwinding problem. In this connection, it should be observed that the total angular momentum of the total system (old double helix plus nucleotide building material, enzymes, etc.) has to remain constant under the entire replication process. This means that, if a certain part of the system rotates around a specific axis, the associated angular momentum has to be compensated for by contributions from other parts moving in an opposite direction.

In discussing various types of replication mechanisms, it is often further clarifying to use a terminology introduced by Delbrück and Stent (1957) which distinguishes between *conservative*, *semiconservative*, and *dispersive* replication. In considering a given parental substance and its daughter duplexes after one or several replications, one says that a process is *conservative* if, among all the duplexes, there is one which is entirely parental and the rest are completely new. A process is *semiconservative* if the parental substance consists of two equivalent parts which are at the first replication divided between the two daughter duplexes so that, after an arbitrary number of replications, there are exactly two duplexes which contain the two parts of the parental substance; these are then said to be the conservative elements of the process. Finally, a process is *dispersive* if the parental substance after several replications has become distributed in small pieces among several of the duplexes. We note that, in this general classification, it is not necessary to make any specific assumption about the character of the parental substance itself. The original Watson–Crick model implies apparently a semiconservative replication mechanism, where the DNA double helix is the parental substance and the strands are the conservative elements which are kept intact in a series of replications.

One of the first attempts to deal with the winding-unwinding problem was made by Delbrück (1955), who assumes that the complementary synthesis proceeds synchronously along the two strands and that, as the synthesis proceeds, the strands break at the growth point at every half-turn of the helix (every fifth link) and the lower terminals of the breaks immediately rejoin the open ends of equal polarity of the new strands. This leads to a *dispersive* replication mechanism.

Another possibility was suggested by Bloch (1955). In his model, there is at the start no unwinding of the sugar-phosphate backbones; after breaking the hydrogen bonds in each base pair, each base is instead turned 180° around the glycoside bonds to open up the "proton-electron pair code" to form a template suitable for synthesis. After formation of two complementary strands on this template, one obtains a four-stranded intermediate compound. The new hydrogen bonds are then broken, the new strands are separated from the old strands, and two standard double helices are formed by rotating all bases 180° around the glycoside bonds to the proper positions for hydrogen bond formation within each complex. The parental helix is then restored and, in addition, one has obtained an identical daughter helix. The use of nucleotides with the base planes rotated 180° around the glycoside linkage has been severely criticized by Crick (1957) for steric reasons, but the model is interesting anyway as an example of a *conservative* replication mechanism.

The so-called Y-model (see Fig. 17) has been studied in greater detail by Levinthal and Crane (1956). They point out that, "by an appropriate combination of rotations of the vertical part and the arms (each on its own axis), all the requirements of the unwinding of the parent and the coiling of the progeny can be satisfied without the Y changing its orientation in space. All that will happen to the Y will be gradual shortening of the vertical part and lengthening of the arms together with a spinning of all three branches." Their calculations show that the procedure "involves requirements of mechanical strength and energy production which are well within those available and that it is the favored type of motion from the standpoint of the viscous drag." The mechanism represents probably the simplest realization of the Watson–Crick model and is a typical example of a *semiconservative* replication procedure.

The rate of the uncoiling of DNA has been calculated by Longuet-Higgins and Zimm (1960) under the assumption that the torque rotating the helix comes from the gain in free energy associated with the entropy increase obtained by letting each strand go over from a rigid structure into a randomly coiled form by utilizing the available freely rotating bonds in the sugar-phosphate backbones. They find that the time required for a separation is of a reasonable order of magnitude, but the model has the difficulty that it does not incorporate the synthesis of the new strands which is simply assumed to occur at some other stage.

We have now given one specific example of each type, but many more have actually been suggested by various authors. A conservative scheme proposed by Stent (1958) will be treated in connection with the RNA formation. For a more complete survey, we will refer to the review articles by Delbrück and Stent (1957) and by Williams (1959).

The multitude of theoretical possibilities for the DNA replication shows that a solution of the problem of the actual mechanism could be found only be refined experiments. An important step towards the solution was taken through the experiments by Meselson and Stahl (1958) who, by means of a heavy trace element (N^{15}), could show that the procedure in *Escherichia coli* is characterized by the fact that "the nitrogen of a DNA molecule is divided equally between two physically continuous subunits; that, following duplication, each daughter molecule receives one of these; and that the subunits are conserved through many duplications." The process is hence definitely *semiconservative.* A similar result was obtained (Taylor *et al.*, 1957; Taylor, 1959) in a study of the chromosome replication of *Vicia faba* (English broad bean) by means of autoradiography, using material prepared with tritium-labeled thymidine.

These experiments gave strong support to the Watson–Crick model, but the character of the intact subunits was still not conclusively shown. Cavaliere and Rosenberg (1961) had experimentally observed double-weight DNA molecules, and they interpreted their results so that the replication process is semiconservative, in agreement with the experiment by Meselson and Stahl, but the parental substance consists of *two* DNA double helices joined by special (biunial) bonds and the conserved unit is a full double helix (cf. Dounce, 1962). Further experiments, however, indicated that the intact subunits in the semiconservative scheme are identical with the single strands in the DNA molecule, in full agreement with the Watson–Crick model, and the previously described Y-model for DNA replication became more and more plausible.

In the meantime Kornberg (1959a, b, 1962) succeeded with his ingenious experiments with biosynthesis of DNA, and the results indicated some complications with respect to the Y-model. Kornberg had pointed out that a nucleotide unit may get attached to a polynucleotide chain by two mechanisms (see Fig. 62), and the experiments indicated that only one of them (B) was actually occurring. The action of DNA polymerase seemed to be such that a lengthening of the chain was achieved only at the end having a free sugar by letting a deoxynucleoside triphosphate attack the hydroxyl group at the C′-3 position of the deoxyribose under elimination of pyrophosphate. This result is in contradiction with the Y-model, which necessarily assumes that the polymerization can occur at both ends of the chain, and the general interpretation was hence that "the Y-type of DNA replication cannot occur" (see, e.g., Setlow and Pollard, 1962).

During the last year there have been some interesting experimental results reported concerning DNA replication by Yoshikawa and Sueoka, by T. Nagata (Nagata, 1963; Nagata *et al.*, 1963), and by J. Cairns, and they all indicate that the DNA molecule replicates as a single piece and that the replication starts from one end only. (See also Butler, 1963.)

Yoshikawa and Sueoka (1963a,b) have studied the genetics of *Bacillus subtilis*, which has a DNA molecule which is easily subject to transformations. The normal form grows in a glucose-salt medium, but by ultraviolet light one may produce mutants which may require a specific growth factor (met, ileu, leu, his, thr, ade, etc.) which then serves as a "marker" for the corresponding locus. Multiple mutants are common, and the order of the markers has been determined by genetic methods; one knows, for instance, that "met" and "ade" are end markers. In a study of DNA molecules having these end markers, Yoshikawa and Sueoka divided the molecules during the replication phase mechanically into small pieces, which were then used for transformation experiments. The result was that, during the replication, there seemed to be twice as many "ade" as "met" markers, indicating that the molecule replicates as a single piece starting from one end, i.e., the "ade" end (see Fig. 60).

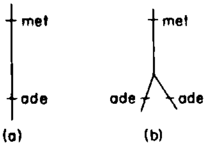


Fig. 60. DNA molecule with end markers (a) in stable form and (b) under replication.

The chromosome of *E. coli* consists essentially of a single piece of two-stranded DNA. In his study of this chromosome by means of autoradiography, Cairns (1963) has used tritium-labeled thymidine for preparing the material and a specific pulse technique (see Fig. 61). This enabled him to

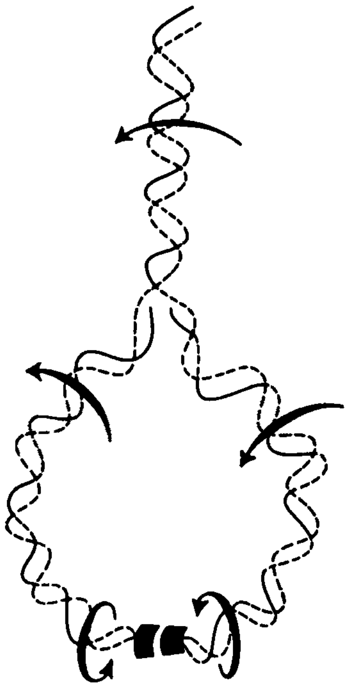


Fig. 61. DNA replication according to Cairns.

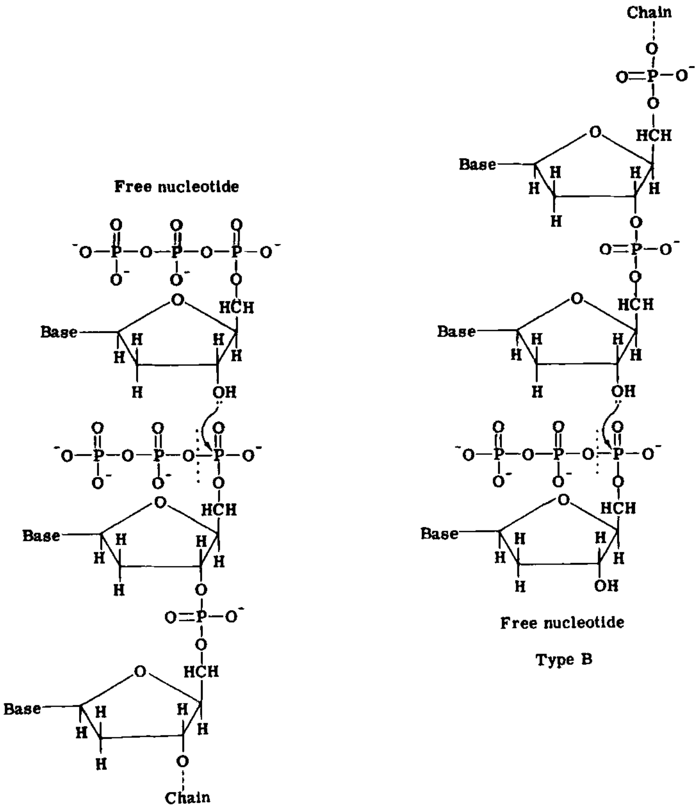
take "pictures" of entire DNA molecules during the various stages of the replication process, and all the results indicate a mechanism of the Y-model type. A somewhat unexpected feature of the pictures was that the ends of the forks were joined by some form of swivel, and that the chromosome of *E. coli* seemed to exist as a circle.

Cairns found further that the DNA molecule was about 700 to 900μ long, and that it hence ought to consist of about 3×10^6 nucleotide base pairs. Since the replication time is around 20 minutes, this implies that the DNA synthesis proceeds with a rate of about 2.5×10^3 nucleotides per second, in good agreement with the data previously found on phage DNA.

Through these experiments, the Y-model seems now firmly established. The synthesis proceeds along both strands, and this means that both Kornberg mechanisms are active (see Fig. 62). This implies further that one has to reconsider the question whether the Kornberg polymerase works differently *in vivo*, or whether there exists actually a second enzyme which is responsible for nucleotide addition of type A. Here one should also raise the question whether the breaking of the energy-rich triphosphate bonds has anything to do with the rotation mechanism, for instance, through ejection of the pyrophosphate into the solution.

B. The Replication Plane

Even if one now knows that the DNA replication follows the Y-model, the details of the actual mechanism are still not available. Here we would like to emphasize the fact that the *steric conditions* during the replication process must be particularly important. According to Section III, the Watson–Crick concept of complementarity depends not only on a fit of the hydrogen bonds but also on a steric factor: a "short" pyrimidine base is supposed to accept only a "long" purine base, and vice versa. Such steric conditions are valid inside the double helix, which is a rather rigid structure stabilized by the histone wrapping in the shallow groove. What happens to this protein chain at replication is not fully known, but it seems likely that it will follow one of the strands during the process, adding stability to its structure. The other strand will probably lose its rigid structure, since the sugar-phosphate backbone contains many single bonds with low rotational barriers (cf. Longuet-Higgins and Zimm, 1960), and this means that the steric factor at the replication is in danger, unless some special precautions are taken. There may be several solutions to this problem, and here we will only point out the existence of a possible "replication plane" with interesting properties.



Type A

Type B

Fig. 62. Outline of the two Kornberg mechanisms for adding a new nucleotide to a polynucleotide chain.

Let us consider the Y-model and the winding-unwinding mechanism (see Fig. 63). It is clear that the two strands of the double helix have to be separated to form a template, but, instead of moving the strands outside the original helix and its protein wrapping as in the original Y-model, we will now investigate what possibilities exist inside the helix itself. It should be observed that, with respect to the axis of the helix, each base covers an angle of about 90°, so that there is actually space for a total of *four helices* without any closer packing of the sugar-phosphate backbones than in the original double helix (Löwdin, 1964a). We will denote the possible helix positions by h_1, h_2, h_3, and h_4.

Considering the normal double helix with the base pairs filling the positions h_1 and h_2, we will start by giving an oversimplified picture. If one could break all hydrogen bonds simultaneously, one could rotate the entire helix h_2 freely around the axis without any distortion into position h_4 with the formation of one new hydrogen bond, and the structure would then have the proton-electron pair code in an open position. The possibility of this rotation to "open the code" has actually been suggested by Crick (1957) as an alternative to Bloch's rotations of the bases around the glycoside bonds. More symmetrically, one could rotate the helices h_1 and h_2 an angle of 90° each in opposite directions to the positions h_4 and h_3, respectively (see Fig. 64). In this "replication form," the two sugar-phosphate backbones are moved more behind the base pairs than before, so that it would be easier to move nucleotides into the helix itself; at the same time, the proton code is still rather well-shielded against outer disturbances which could change the genetic message. In this oversimplified form, the entire double helix serves as a template and, since each plane contains one purine and one pyrimidine base, the steric conditions are strongly emphasized. Each base will now only accept a partner which will fit both the hydrogen bond code and the helix's geometry (see Fig. 65). After conclusion of the synthesis according to the complementarity principle, one would in this oversimplified form have a *quadruple helix* with the two pairs h_1—h_2 and h_3—h_4 joined by the conventional hydrogen bonds (four to six bonds in total), whereas the two halves would be joined by at least two temporary hydrogen bonds. [For a related but different mechanism, see Fong (1964).]

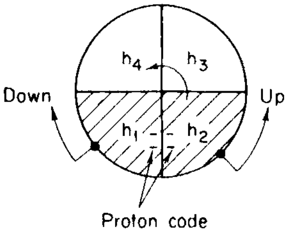


Fig. 63. Geometry of a base plane in DNA.

The actual replication scheme will now be described in close accordance with the Y-model, and the quadruple helix will be limited to consist of a single *replication plane* situated at the branching point of the two forks (see Fig. 66). It is important to observe that the process will be *semiconservative*, and that the single strands represent the intact units.

It is assumed that, during the replication, the replication plane moves from one end of the original helix to the other leaving two separated daughter helices behind. In such a procedure no large initial energy is required, since the process involves a simultaneous synthesis and decomposition. In each replication plane, there are five to seven new hydrogen bonds formed when the complementary nucleotides are brought in, and the energy released may be used to break two or three hydrogen bonds in the primer helix and the two "temporary" hydrogen bonds in the previous replication plane. At a given time, only one plane of the original double helix will be in "replication form" (see Fig. 64), whereas the main

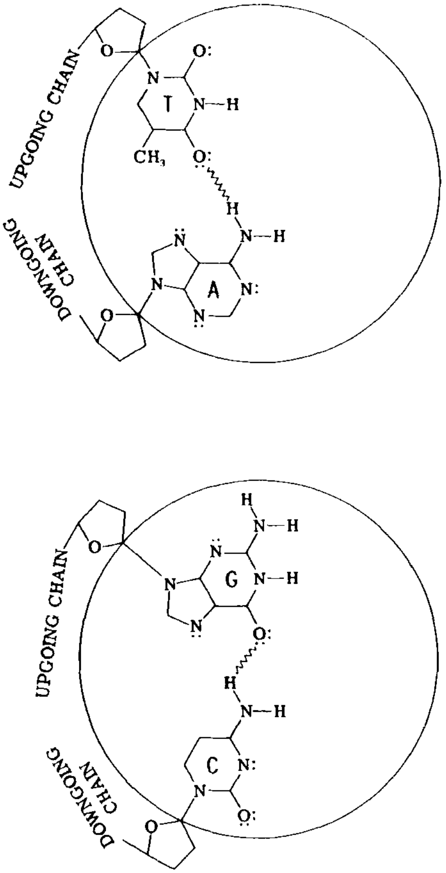


Fig. 64. The two nucleotide base pairs in the suggested "replication plane" with the proton-electron pair codes open for the addition of a pair of complementary bases. The symbol /\/\/\ indicates a temporary hydrogen bond.

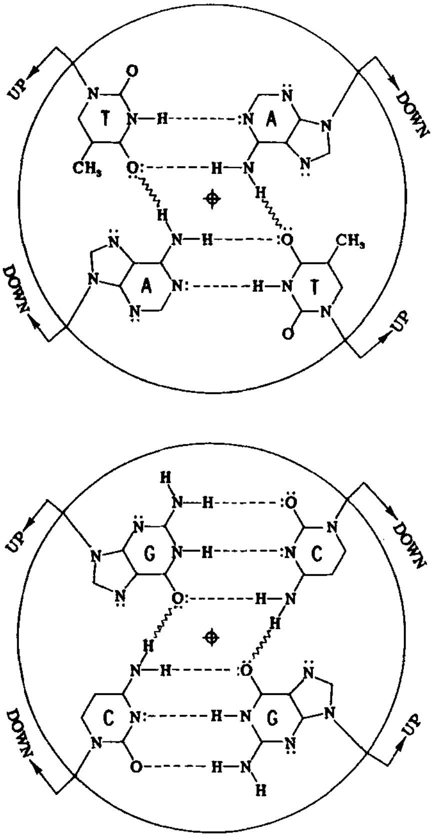


Fig. 65. The four bases in the suggested "replication plane" after the synthesis. The two base pairs are afterwards separated by breaking the temporary hydrogen bonds indicated by the symbol /\/\/\. These hydrogen bonds may be of importance in preventing a single tautomeric form to enter the replication plane. Note that the electron lone pairs on the N-7 position of the purine bases may participate in weak hydrogen bonds which are not indicated on the figure.

part of the remaining helix will be in normal form (see Fig. 66). In the short region between the two forms, the helix is in an intermediate structure with the sugar-phosphate backbones slightly deformed corresponding to an "unwinding" of the helix. We note that, in this picture, the synthesis in the replication plane steers both the unwinding of the old helix and the winding of the two daughter helices. It seems further very likely that the building zone is approximately at rest, so that the entire picture is in agreement with Levinthal and Crane (1956).

The replication plane offers the strict steric conditions which are necessary for obtaining the Watson–Crick type of complementarity. At the same time, there are two additional temporary hydrogen bonds which may be of importance in "purifying" the nucleotide material brought into the synthesis, in order to avoid "incorporation errors" and to secure the correct base pairing.

We note that in the middle of the replication plane there is a characteristic parallelogram which involves two amino groups and two ketogroups situated in opposite corners. One has here a double pairing between these groups in the original sense of Chargaff (1955, 1957, 1963). This double pairing is of importance since it may prevent the incorporation of a single

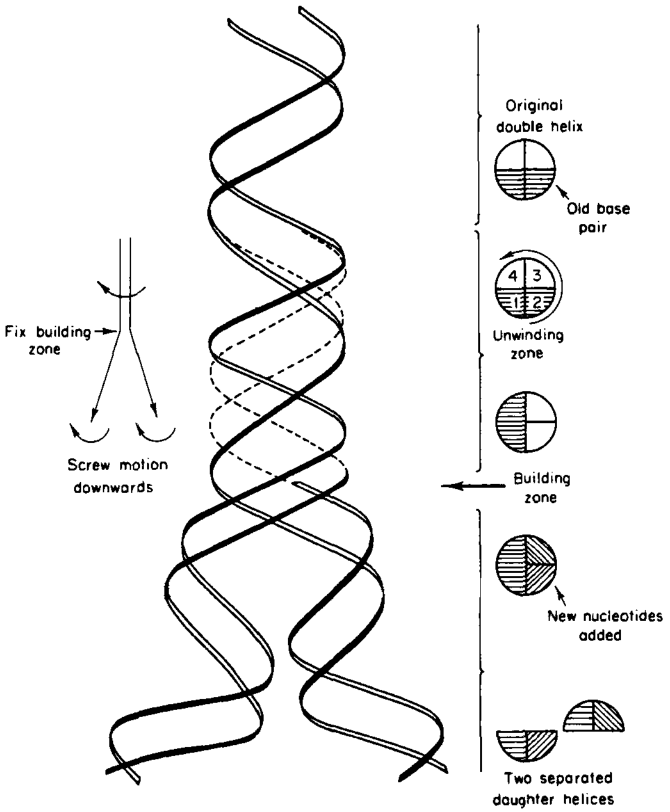
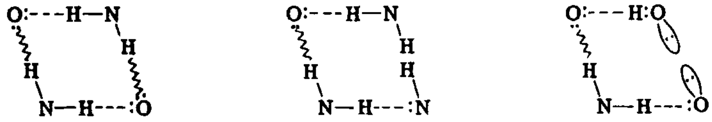


Fig. 66. Formation of the two daughter helices in the suggested mechanism of DNA replication.

tautomeric form of a base from the nucleotide material in the environment (Löwdin, 1964c).

As an illustration, we will study the A—T replication plane in its normal form compared with the two planes involving a single tautomeric base which could be incorporated according to the rules discussed in Section III, C. Considering only the four key groups around the center, one obtains the diagrams shown in Fig. 67. It seems as if the incorporation of a single tautomer would be prevented by the fact that either two protons or two electron pairs would meet, whereas the incorporation of the correct base is favored by the fact that an additional temporary hydrogen bond is formed. A similar reasoning holds for the G—C plane.



(a) (b) (c)

Fig. 67. The four key groups around the center of the replication plane in normal replication and in the case of single tautomeric incorporation errors: (a) TA—AT plane, (b) TA—AC* plane, (c) TG*—AT plane.

It should be mentioned that the electron lone pair on the N-7 position of the purine ring may also play a role in this connection. In the A—T replication plane it may form a weak temporary hydrogen bond with the protons in the CH_3 group of thymine and in the G—C plane it may compete with the electron pair in the keto group (O:) in forming a temporary hydrogen bond with one of the protons in the amino group (NH_2). These hydrogen bonds are probably rather weak energetically, but they may still play a certain steric role in the properties of the replication plane.

Actually a "purification" procedure in the DNA synthesis seems to be needed in order to prevent excess incorporation errors, since the tautomeric forms of the nucleotide bases do not seem to be as "rare" as originally anticipated. Measurements have shown that, for some of the pyrimidine bases, the equilibrium constants are comparatively high. Some empirical data are listed below:

methylcytosine:	$K = 10^{-4.7}$	(Kenner *et al.*, 1955)
uracil:	$K = 10^{-3.3}$	(Katrizky and Waring, 1962)
bromuracil:	$K = 10^{-1.7}$	(Katrizky and Waring, 1962)

This means that, if the nucleotide material in the environment of a replicating DNA molecule is at least approximately in tautomeric equilibrium, it would be easy to get quite a few tautomeric errors incorporated, and the mutation rate would then be correspondingly high. Actual studies of mutation rates (Freese, 1962) indicate that the probability for a spontaneous change of a base pair lies in the range 10^{-8} to 10^{-11} per generation, and this figure may be even lower. To solve this dilemma, some authors have assumed the existence of a special enzyme which could take care of the "purification" of the nutrition material (Koch, 1963). It seems now as if at least part of this purification could be handled by the "replication plane."

It should finally be observed that the mechanism of the replication plane does not prevent the incorporation of "tautomeric base pairs." The probability for such an event is regulated by the product $K \times K'$, where K and K' are the equilibrium constants for a complementary pair of bases.

Unfortunately, the equilibrium constants of the purine bases have not yet been accurately measured, but it does not seem unreasonable to assume that the products KK' will be in the range of 10^{-8} to 10^{-10}, in better agreement with Freese. Some of the base analogs, like bromuracil, may be exceptions which explain their mutagenic character.

C. Formation of RNA

The hereditary substance should have not only the property of being self-replicating but also the ability of transcribing the genetic message to the cell and the organism. As pointed out in Section III, there seem to be good reasons for believing that the genetic code is transferred from the DNA molecule to the ribosomes in the cytoplasm, where the actual protein synthesis takes place, by means of a special polynucleotide called *messenger RNA*.

If the mechanism of the DNA replication is rather uncertain, the actual procedure involved in the formation of this messenger RNA is still less known. Experimental evidence (Volkin and Astrachan, 1956; Belozersky and Spirin, 1958; Volkin *et al.*, 1958; Ycas, 1959; Ycas and Vincent, 1960; Volkin, 1960; Nomuar *et al.*, 1960; Stevens, 1961) indicates, however, that the base ratios of RNA are closely analogous to those found in the DNA regulating the formation, and it seems hence natural to assume that DNA in some way acts as a template. The simplest assumption is that the transcription process occurs very much the same way as the replication process: the genetic code is temporarily opened, and one of the strands offers a template for the synthesis of a single-stranded messenger RNA molecule.

In this section, we will also briefly review some other attempts to describe the transcription mechanism without opening the genetic code.

The double helix of DNA has a shallow groove and a deep groove which are both rich in protons and lone electron pairs. Today, it is believed that the shallow groove is occupied by a helical protein (histone) which helps in stabilizing the entire molecule (see, e.g., Wilkins, 1957), whereas the role of the deep groove is not yet fully known. It was observed by Stent (1958) that the deep groove contains a proton-electron pair code which may serve as a template for a third helix which would fit into this place. The hydrogen bond formation suggested by Stent is given in Fig. 68.

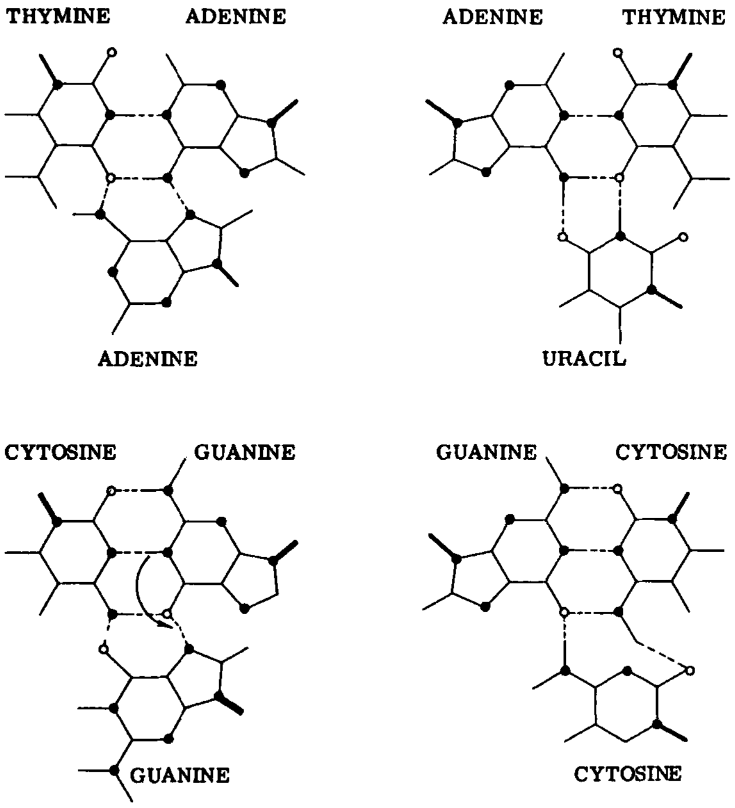


Fig. 68. Transcription of the genetic code, according to Stent (1958).

In this figure, filled and open circles represent nitrogen and oxygen atoms, respectively; the little arrow indicates a tautomeric shift. If one actually builds a stereomodel of this triple helix, one finds that, in order to obtain a connecting sugar-phosphate backbone for the third helix, it is necessary in two of the cases to rotate the entire side chain in the nucleotide 180° around the glucoside bond. However, according to Crick (1957), this form of the nucleotides seems sterically to be rather unprobable.

Of crucial importance in the problem of the DNA-RNA transcription is the question of the relation between the two quotients:

$$\left(\frac{G+C}{A+U}\right)_{\text{RNA}} \qquad \left(\frac{G+C}{A+T}\right)_{\text{DNA}}$$

Two more triple helices have been suggested by Zubay (1958a,b, 1962), and they have the transcription characteristics shown in the tabulation below.

Model A	Model B
$TA \rightarrow C$	$TA \rightarrow A$
$AT \rightarrow G$	$AT \rightarrow U$
$CG \rightarrow U$	$CG \rightarrow G$
$GC \rightarrow A$	$GC \rightarrow C$

They are described in greater detail in Fig. 69. Both are within certain margins sterically reasonable. Since some of Zubay's experiments (1958a,b) indicated that the values of the above-mentioned base ratios were the inverse of each other, Zubay favored "model A" and wanted "model B" eliminated. On the basis of other experiments, one may today feel more inclined to favor "model B," and a more complete discussion of this problem will be given below.

Experiments with synthetic polynucleotides (Rich, 1958a, b, 1959a, b, c, 1960) have shown the existence of three-stranded helical molecules, and polyinosic acid with three hypoxanthine bases in the plane is particularly remarkable because of the symmetry in the hydrogen bonding between the constituents. Study of such three-stranded complexes is certainly of value in this connection.

An important experiment concerning the nature of the DNA-RNA transcription in the bacteriophage T2 has been reported by Hall and Spiegelman (1961). They have found that the RNA molecules synthesized in the bacteriophage-infected cells of *E. coli* have the ability to form a well-defined complex with denaturated DNA of the virus. From the

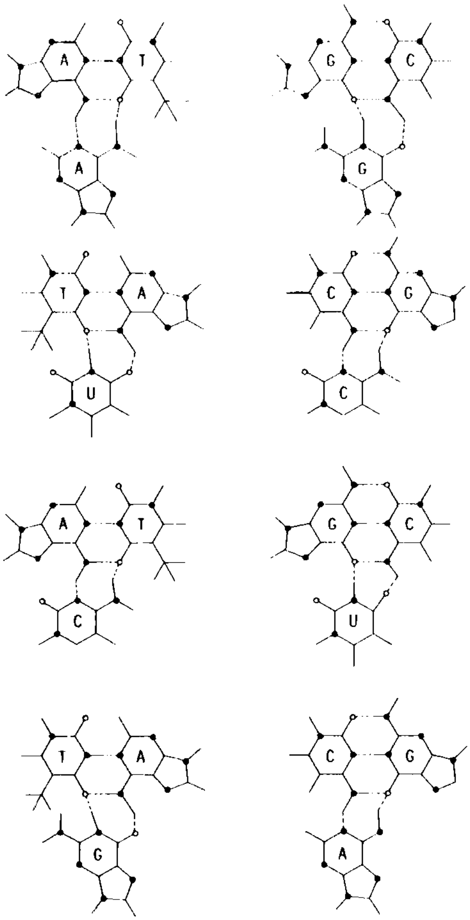


Fig. 69. Combination of a DNA base pair and a third nucleotide base in the deep groove of DNA according to Zubay (1958). Filled and open circles represent nitrogen and oxygene atoms, respectively. Note that the lowest base in each figure represents the RNA component.

formation of this hybrid, they conclude that there must exist a perfect, or near-perfect complementarity between the nucleotide sequences of T2-DNA and RNA in the sense of Watson and Crick (1953): "Demonstration of sequence complementarity between homologous DNA and RNA is happily consistent with an attractively simple mechanism of informational RNA synthesis in which a single strand of DNA acts as a template for the polymerization of a complementary RNA strand." Let us now study the conditions for such a mechanism in greater detail.

In investigating the normal base pairs (see Fig. 16) one finds immediately that the proton-electron pair code exposed by one of the bases also occurs in the "deep groove" (see Fig. 70). The distance between the proton and the lone pair concerned is, however, considerably longer than in the base, and also the angular structure needs improvement. The situation becomes much more ideal for "transcription" if one can rotate the base pairs around the axis somewhat towards each other, say 15–30°, to get a better analogy with the structure of polyinosic acid mentioned above (cf. Rich, 1959a, b, c). It should be observed that this rotation of the two helices should be possible without breaking the bonds with the histone wrapping in the shallow groove, since the protein chain contains some extra

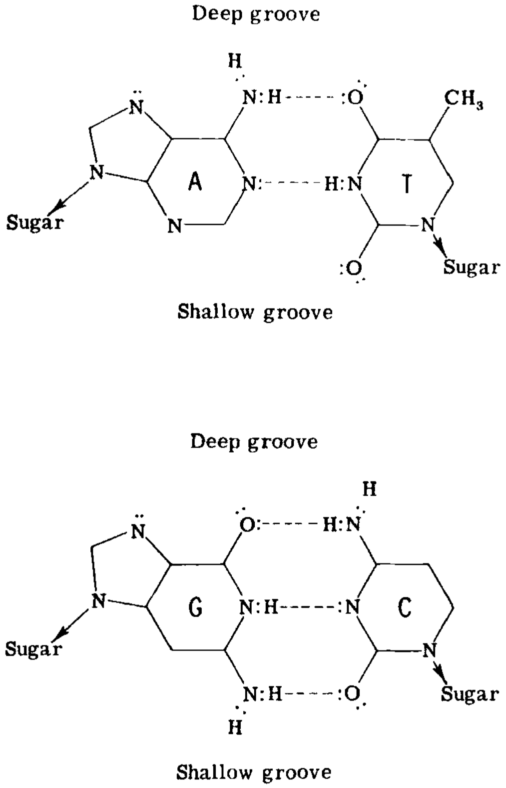


Fig. 70. The proton-electronpair code of the base pairs in the deep groove of DNA.

folds. This operation brings the double helix into a "transcription form," and a third helix can now easily be added in the deep groove with the base determined by the hydrogen bonding as indicated in Fig. 78.

A few words should perhaps be added about the geometry of the third helix. Since the pyrimidine bases are considerably "shorter" than the purine bases, it turns out that, in the attached pentose groups, the 3′ position of the former is approximately equivalent to the 2′ position of the latter. One may wonder whether the two positions should be also chemically equivalent, i.e., carry an hydroxyl group, and whether this may be an explanation

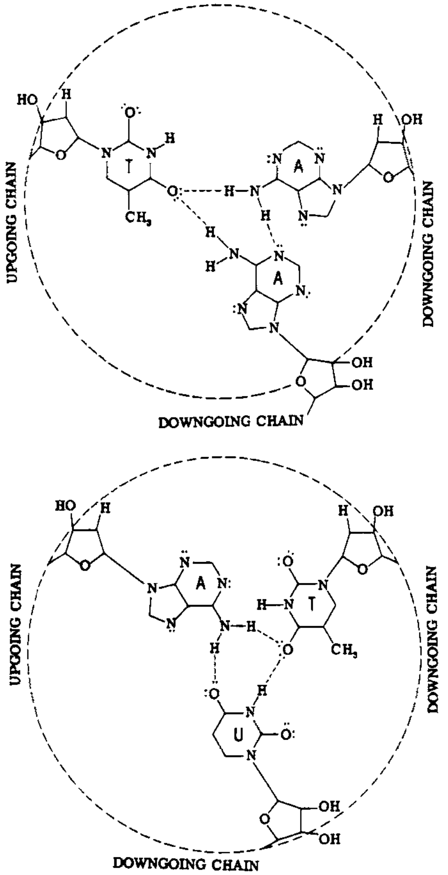


Fig. 71. Combination of DNA base pairs and a third nucleotide base in the deep groove of DNA in "transcription form."

why the third helix should have ribose character and be RNA. It is well-known that, in RNA as well as in DNA, one has so far found only C_2'—C_5' linkage (see, e.g., Davidson, 1960a, b; Chantrenne, 1961), but it may be worthwhile checking whether messenger RNA would show any C_2'—C_5' linkage as well as C_3'—C_5' linkage. The problem of the meaning of the difference in the character of the 2′ position between RNA and DNA is still far from being solved.

Actually there is a great deal of equivalence between the "transcription planes" in Figs. 71 and 72 and the "replication planes" in Fig. 65. However, the steric conditions of the former are not as strict as in the latter, and under these circumstances one would expect that sometimes in transcription

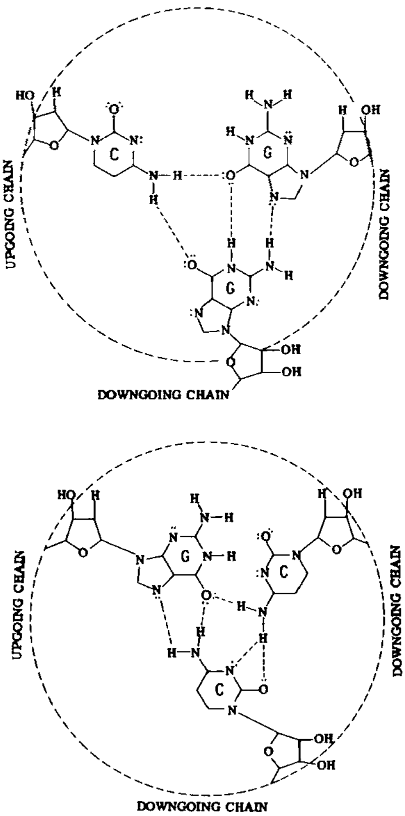


Fig. 72. Combination of DNA base pairs and a third nucleotide base in the deep groove of DNA in "transcription form."

a pyrimidine base might be replaced by a code-equivalent purine base, so that uracil might be replaced by guanine and cytosine by adenine. A study of the multiple hydrogen bonding in the transcription plane shows, however, that the groups in the 2 positions add extra specificity to the transcription procedure, so that one obtains rather exact "reading rules" A → A, T → U, G → G, C → C. It should still be observed that the experiment by Hall and Spiegelman (1961) does not necessarily prove *complete* complementarity in the sense of Watson and Crick; it is probably sufficient for the formation of the hybrid that the strands involved are complementary with respect to the hydrogen bonding alone.

It is evident from Figs. 71 and 72 that only one of the strands of the double helix DNA will be "read" in the transcription procedure, as the formation of RNA proceeds along the axis. As in the replication of DNA the energy released in forming new hydrogen bonds will be used to break previously formed bonds, which means that, as the procedure goes on, the newly formed RNA strand will become free from the double helix. The unwinding of the new strand perhaps follows a mechanism similar to the one in the DNA replication, and the torque may come from the splitting of the nucleoside triphosphates. It should be observed that, if only one strand of DNA has been faithfully read, one could understand the similarity between the ratios (G + C):(A + T) for DNA and (G + C):(A + U) for RNA.

If only one strand of DNA is read in transcription, one cannot understand Chargaff's second rule A + C = G + U for RNA. Instead one has to assume that the transcription procedure goes along the DNA molecule the distance of a "gene" and returns to the starting point. This means that it first reads one strand and then goes back reading the other strand, and when the process has reached the starting point again there has been synthesized a single-strand molecule with a "loop" connecting the two RNA halves coming from the two DNA strands. If there were no degeneracy, the base composition of the two halves of RNA would be an exact copy of the base composition of DNA and, according to Chargaff's first law, one would have A = U and C = G. This RNA molecule would further be characterized by the fact that it would be a single-strand molecule which would be identical with its own complement. However, if code-equivalent bases under certain conditions may replace each other, so that A replaces C and G replaces U, one obtains instead A + C = G + U, i.e., the law Chargaff has found characteristic for RNA and expressed in the form "6-amino = 6-keto."

This type of reading procedure would also explain the remarkable tendency of the RNA molecule to fold back on itself without necessarily giving rise to a double helix (which may be prevented partly by the steric disturbances introduced through the above-mentioned degeneracy). By using the arguments of Hall and Spiegelman, this tendency may be taken as an indication for the existence of at least code complementarity between the two halves of the molecule.

A fundamental problem discussed frequently today is the question whether only one or both of the DNA strands are "read" in the formation of messenger RNA and the associated protein synthesis. There seems to be some experimental evidence as to bacteriophage DNA that both strands are read in "replication," since changes in one strand give rise to both wild-type and mutant progeny (Pratt and Stent, 1959; Tessman, 1959; Vielmetter and Wieder, 1959). It seems to be more difficult to show what happens in transcription, but some *in vitro* experiments indicate that both strands of DNA are actually copied (Geiduschek *et al.*, 1962; Hayashi *et al.*, 1963). However, it seems as if most workers in this field now have the opinion that *in vivo* only one strand of DNA is read, and a certain amount of experimental evidence has been found (see, e.g., Guild and Robinson, 1963). This implies also that Chargaff's second rule may not hold for messenger RNA, even if it holds for the main bulk of the other RNA molecules. Special biological functioning of the two DNA strands in transcription has been suggested by Paigen (1962).

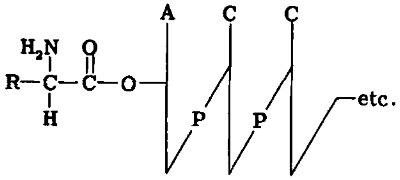
Part of our discussion has been based on the idea of the existence of multiple helices (Rich) and on the observation that there exists a copy of the genetic code in the deep groove of DNA which may be used for transcription (Stent, Zubay). It is now easy to understand the relation between the coding mechanism outlined in Figs. 71 and 72 and the one previously suggested by Zubay (Fig. 69). His "model B" is identical with the transcription we obtain in the reading of the "second" strand, whereas his "model A" represents the same transcription changed to code-equivalent bases, so that $A \leftrightarrow C$ and $G \leftrightarrow U$.

Several other mechanisms for the transcription process have been suggested, e.g., the opening and reading of the ordinary genetic code or the formation of a "stencil RNA" (Ochoa, 1962). Many authors consider the former to be the most plausible one, but in order to make a definite decision more experimental results are certainly necessary.

D. Coding Problem: Recent Aspects

As pointed out in Section III, it is believed that the protein synthesis in the ribosomes is regulated by an "adapter mechanism" as suggested by

Crick (1958) and by Hoagland (1959). Each one of the twenty amino acids is attached to an adapter in the form of a comparatively small molecule of *soluble* RNA (sRNA) which carries the amino acid to its correct place on the messenger RNA by means of complementarity between a small nucleotide part of the adapter and the template. This process is the subject of intense experimental study, and it has been found, for example, that the amino acid is always attached to the adapter in the same way (Lipmann *et al.*, 1959), namely through a nucleotide terminal group of the type ACC:



It is not yet completely clear whether the amino acid linkage goes to the 2′ or 3′ position of the pentose group in adenosine. The adapter contains, in addition to the nucleotide triplet and the sequence characteristic for the adapter and its amino acid, quite a few other constituents which are now being investigated (Zubay, 1963; Langridge and Gomatos, 1963).

Let us now return to the coding problem treated in Section III. According to Crick *et al.* (1961) the code ought to be a nonoverlapping triplet code without commas which is read in triplets from a specific starting point. The overlapping codes were eliminated since they led to correlations between the dipeptides which do not exist in nature, and the "comma-less" codes had to be abondoned because of the existence of sense triplets of the form UUU, as shown in the experiments by Nirenberg and Matthaei (1961). The code is certainly degenerate, and the question is now how much can be said about the nature of the degeneracy. In this connection, one can follow part of the arguments given by Gamow provided that one changes the nature of the code to be nonoverlapping instead of overlapping.

Let us start with the *permutation-invariant code*, in which the message is independent of the order of the three bases involved in the triplet. Since this code is equivalent with Gamow's "triangular code," it contains exactly twenty independent elements and each amino acid would have its unique triplet of letters. However, since the experiments (Nirenberg and Matthaei, 1961) indicate that the coding letters for both leucine and valine may be U, U, and G, and for both glycine and tryptophan may be U, G, and G, this code should be excluded.

The *direction-invariant code* offers the next possibility. In this code, the message is independent of the direction the triplet is read, so that $B_1B_2B_3$ and $B_3B_2B_1$ are equivalent. The number of independent combinations is given below:

	BBB		$B_1B_2B_1$		$B_1B_2B_3$	
number of combinations	4	+	12	+	24	=40

The total number is forty, and this may be reduced by a factor of two by assuming that the *middle position distinguishes only between two possibilities*, so that the code is degenerate in this position with respect to complementarity (which leads to a nonoverlapping code equivalent with Gamow's "diamond code") or with respect to code complementarity, code similarity, and so on.

A somewhat different code is based on the assumption that, if both strands of the DNA double helix are read in their specific directions, they would contain the same message with respect to the protein synthesis. This means that the triplet $B_1B_2B_3$ should be equivalent with the triplet $B_3'B_2'B_1'$, where the symbol B' means the complementary basis to B. It leads to the diagram:

	BBB		$B_1B_2B_1$		$B_1B_2B_3$	
number of combinations	2	+	6	+	24	=32

with thirty-two combinations. This *strand-invariant code* is easily reduced to only twenty independent elements by an additional assumption about a specific degeneracy. However, in the present situation, it is hardly worthwhile to go through all the mathematical possibilities, since "strand invariance" does not seem to be a fundamental concept if only one strand of DNA is read in transcription. It would, of course, be particularly simple if the degeneracy of the code was connected with a symmetry property, but this is not necessarily the case.

In spite of all its complexity, the coding problem seems now to be close to its solution. In Section III, F, we briefly mentioned the new approach to this problem introduced in 1961 by Nirenberg and Matthaei (Nirenberg *et al.*, 1962; Singer *et al.*, 1963) who succeeded in using synthetic polynucleotides as templates in the biosynthesis of polypeptides. This technique has been further developed by Ochoa and his group (Speyer *et al.*, 1962) and by other authors (Wood and Berg, 1962). It has been supplemented by a method observing the replacement of tobacco mosaic virus (TMV) proteins when the TMV—RNA is treated by nitrous acid, which acts as a strong mutagen (Wittmann, 1961; Tsugita and Fraenkel-Conrat, 1960, 1962; Quastler and Zubay, 1963). The development in this field goes so fast that the results very quickly become obsolete, and, for a survey of the present situation, we would like to refer to some recent review articles (Gavrilov and Zograf, 1963; Burma, 1962). A summary of the results is given in Table VI.

TABLE VI

THE TRIPLET CODES PRESENTLY ASSIGNED TO AMINO ACIDS BY NIRENBERG AND MATTHAEI (N) AND BY OCHOA (O)

Amino acid	Triplet	Observer	Amino acid	Triplet	Observer
Gly	GUG	NO	Asp NH_2	ACA	NO
	GCG	O		AUA	O
	GAG	O		AUC	O
Ala	CCG	NO	Glu	AAG	NO
	CUG	O		AUG	NO
				UGC	NO
	CAG	O	Glu NH_2	AAC	NO
Val	UUG	NO		UGC	O[a]
Leu	UAU	NO	Lys	AAA	NO
	UGU	NO		AGA	N
	UCU	NO		AAU	NO
I'leu	AUU	NO	Arg	CCG	NO
	AAU	O		CUG	O
Ser	CUU	NO		AAG	O
	CCU	NO	Phe	UUU	NO
	CAG	O		UCU	O
	CUG	N	Tyr	AUU	N
Thr	ACC	N	Try	GGU	NO
	AUC	O	Hist	ACC	NO
	AAC	NO		AUC	O
	UCC	O	Prol	CCC	NO
CySH	UGG	NO		CUC	NO
Met	UAG	NO		CAC	NO
	UGG	NO		CGC	N
Asp	ACG	O			
	AUG	O			

[a] Added from replacement data.

The question of the degeneracy of the code and possible symmetry properties have been considered by several authors (Woese, 1962; Eck, 1963), but a definite solution has not yet been obtained. The empirical data and their interpretation have been critically evaluated by Crick (1962) who also gives an excellent survey of the entire problem.

In conclusion, we observe that we have here discussed the replication mechanism, the transcription procedure, and the genetic code as if they all would be of universal nature. The existence of single-stranded DNA in the bacteriophage ϕX174 (Sinsheimer, 1959), the direct replication of the RNA of tobacco mosaic virus, etc., indicate that this can hardly be the case and that there must be many exceptions to the general pattern. One would perhaps anticipate that the single-stranded forms have a simpler replication pattern, but, among other things, it has now turned out that the bacteriophage ϕX174 has a rather remarkable way of replication involving a two-stranded DNA (Sinsheimer, 1962; Sinsheimer *et al.*, 1963; Hayashi *et al.*, 1963).

VII. Problem of Mutations

A. Spontaneous Mutations

In this section, we will discuss some aspects of the biological functioning of DNA in view of the Watson–Crick model. Let us start with the problem of the spontaneous mutations and discuss them on the basis of the tautomeric mechanism. According to the ideas developed in this paper, genetic errors could be spontaneously introduced through the following mechanisms:

(1) incorporation errors depending on the fact that a pair of tautomeric bases has been introduced into the replication plane;
(2) inherent code errors depending on spontaneous proton transfer in the hydrogen bonds of the base pairs in DNA.

The incorporation errors have been treated in Section VI, B and depend on the product $K \times K'$ of the equilibrium constants for a pair of complementary bases in the environment of the replicating DNA molecule. Unfortunately, there are not yet sufficiently accurate measurements available to make a definite statement concerning the order of magnitude of this effect.

In the quantum-mechanical study of the properties of the hydrogen bond in Section V, it was shown that there is a very small but still finite probability for simultaneous "proton tunneling" through the potential barrier in the middle of the bond leading to the transitions

$$\mathrm{A{-}T} \rightarrow \mathrm{A^*{-}T^*}, \qquad \mathrm{G{-}C} \rightarrow \mathrm{G^*{-}C^*},$$

showing that normal base pairs may be changed into *pairs* of tautomeric bases. Irrespective of whether the tautomeric pair is established by an incorporation error or proton tunneling, one can expect the replication diagrams shown in Fig. 73. Hence these effects both lead to the following change of base pairs:

$$\mathrm{A—T} \longrightarrow \mathrm{G—C}$$

$$\mathrm{G—C} \longrightarrow \mathrm{A—T}$$

where a base goes over into another base of the same type, i.e., a purine into a purine and a pyrimidine into a pyrimidine. Mutations of this type have been called "transitions" and are characterized by the fact that they are reversible.

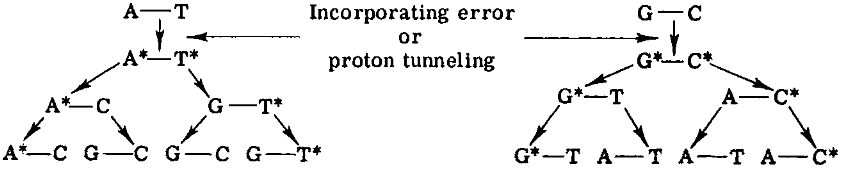


Fig. 73. Replication scheme effected by proton tunneling or a double incorporation error.

It should be observed, however, that these replication schemes are valid only if one assumes that the intermediate base pairs of type $G^*—T$, $A—C^*$, ..., etc., are stable against proton exchange. It has previously been pointed out that the forms $G^*—T$ and $A—C^*$ have absolutely lower π-electron energies than the forms $G—T^*$ and $A^*—C$, respectively, and preliminary studies indicate that the same applies also to the total energies. This means that the forms $G—T^*$ and $A^*—C$ are unstable with respect to proton exchange and quickly go over into the forms $G^*—T$ and $A—C^*$, respectively. This causes further a change in the replication scheme in the left-hand part of Fig. 73 which instead takes the form illustrated in Fig. 74, showing a stabilization of the $A—T$ pair.

The problem of spontaneous mutations is complicated by the fact that the order of magnitude of both the effects involved is still essentially unknown. Measurements in the future will probably give us the values of the equilibrium constants, K and K', for the tautomeric pairs, but one still does not know whether the nutrition material around the DNA molecule during replication is close to tautomeric equilibrium or not. With respect to proton tunneling within DNA, our preliminary study has shown that all the hydrogen bonds seem to have enough asymmetry to render them an immense stability in a situation of thermal equilibrium at ordinary temperatures, except possibly for the middle bond of the G—C pair. The difficulty lies in the fact that even the mobile electron distribution has to be evaluated with a great deal of accuracy, and that an error of $\Delta q = 0.01$ in a charge order may correspond to a factor of ten in the proton transfer rates. A recent comparison (Miller and Lykos, 1962) between the Hückel charge orders of the purine molecule and those obtained by a semiempirical self-consistent molecular orbital calculation indicated changes in Δq in the range between 0.004 and 0.30, and similar results have been obtained by other authors (Veillard and Pullman, 1963; Rein and Ladik, 1964). In this situation, one should also remember that one has still no criterion as to the reliability of the results.

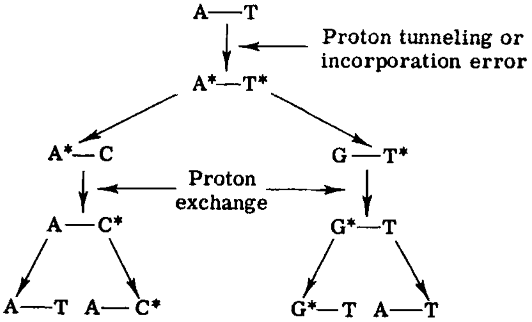


Fig. 74. Replication scheme effected by a second proton exchange leading to a stabilization of the A—T base pair.

It is difficult also to give a reliable figure for the mutation rates based on genetic experiments, but Freese (1962) has estimated that the probability for a spontaneous change of a base pair should lie in the range 10^{-8} to 10^{-11} per generation. In this connection, one has, of course, to observe that not necessarily every change in a base pair corresponds to a real mutation.

In discussing the importance of tunneling errors, one should observe that the time interval between two DNA replications may vary from a few minutes, as, e.g., 20 minutes ($\approx 10^3$ sec) for *E. coli*, to many years for certain types of cells. The discussion in Section V indicates that, for the double-well potentials occurring in DNA, the proton tunneling during the time between two replications may be of a correct order of magnitude to give a reasonable explanation of the occurrence of the phenomenon of spontaneous mutations. Actually both incorporation and tunneling errors may be of importance and, for the moment, one has no possibility to decide whether one or another is dominating. The main point in our study is perhaps that both processes are of such a nature that they lead to rare events and do not destroy the normal stability of the genetic code.

The transcription process may, of course, also influence the stability of the genetic code, and the mutation rate could hence also depend on how many times a code has been transcribed between two DNA replications. Since little is known about the transcription mechanism, one cannot draw any definite conclusions, but one certainly has here an interesting field of research. In conclusion it should also be remembered that, in reality, it is very difficult to distinguish between spontaneous and induced mutations, particularly if one considers the effect of small perturbations, for instance, the fluctuations of the charge distributions.

B. Induced Mutations

In addition to the spontaneous mutations, there are mutations induced by certain chemicals (mutagens), by radiation, and by other effects (see, e.g., Schull, 1962). The simplest mutagens are those which cause a major change in the bases themselves connected with a change in the proton-electron pair code, e.g., nitrous acid which causes an oxidative deamination of the bases (see Fig. 75). Since the base code has been changed, the process will necessarily induce mutations as, for instance, shown in Fig. 76.

Other types of mutagens may just shift the π electron cloud around positions 1 and 6 enough to induce a proton transfer within a base and lead to a tautomeric form. Mutations may also be caused by adding base analogs (like 5-bromuracil) in the synthesis of DNA. If such a base analog has two tautomeric forms, the base analog may be accepted in one form but, after proton transfer, it may replicate in the other form leading to a genetic error. This process may be of essential importance in destroying the genetic code in cells or organisms which are for some purpose not desired. The remarkable success of Kaufman (1963) and Kaufman *et al.* (1962a, b, c) in treating the virus disease *Herpes simplex* in the eye by means of a 5-iododeoxyuridine is a good example of this new approach in chemotherapy.

It has previously been pointed out that proton transfers of the second type, where A—*T* goes over into A^+—T^- and A^-—T^+, may be caused by inducing unequal charges in the two bases. This can be achieved by adding or subtracting electrons or protons. Since the mutagenic ions do not combine with any one of the normal bases, they may in certain cases cause

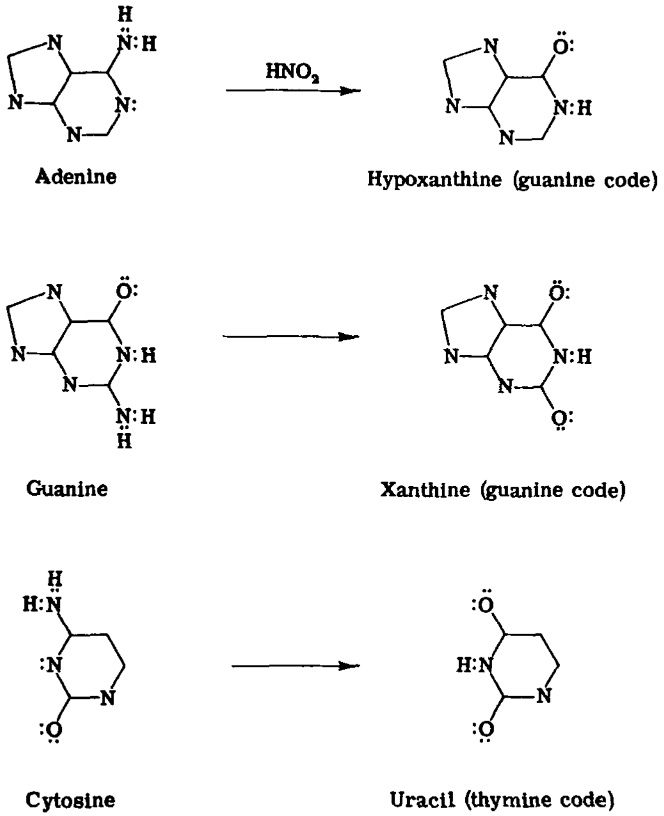


Fig. 75. Effect of nitrous acid on the nucleotide bases.

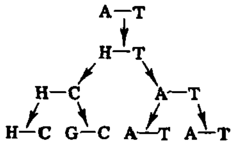


Fig. 76. Replication scheme effected by the influence of nitrous acid. H = hypoxanthine.

deletions and hence irreversible mutations. According to some authors (Freese, 1959a, b; Sager and Ryan, 1962), irreversible mutations would depend essentially on "transversions," i.e., changes of base pairs where a purine base is replaced by a pyrimidine base, and vice versa. The essential importance of deletions and additions in this connection has been stressed by Crick and his group (Brenner *et al.*, 1961).

Certain chemical compounds (often in the form of conjugated systems) may be supposed to enter into donor-acceptor reactions with the normal base pair, leading to an unbalance of the charge distribution, but it is observed that only very special types have a mutagenic effect. This may be caused by the fact that the base pairs are normally extremely well-shielded so that, in order for the mutagen to interact with them, it is necessary that it can reach them, for instance by being attached to the special proteins which catalyze the DNA replication itself. Attention should also be devoted to the possibility for protons (or small positive groups) moving in the environments of the base pairs to get attached to some of the extra lone pairs existing, since this will give an extra charge to the base involved. Another important point is the removal of an electron through ionizing radiation, which may have a more or less direct influence. The effect of electronic excitation of the $\pi \rightarrow \pi^*$ type within a base pair has also recently been emphasized (Rein and Ladik, 1964). The effect of electron or proton "bombardment" has so far not been sufficiently studied.

In conclusion, it should be mentioned that one may expect induced mutations if the electrostatic double-well potentials in the hydrogen bonds are disturbed by an *outer electric field*. It seems hence worthwhile to carry out experiments involving the growing of cultures in specially applied electric fields (Bozoky *et al.*, 1963) or on plastic or other materials which have a large ability to keep high electrostatic charges for a very long time in the form of double layers.

C. Theory of Evolution

It is evident that the mutations must play a most important role in the theory of evolution. The first polynucleotide with the property of self-replication has been called the first "living thing" on earth (Horowitz, 1945; Oparin, 1953, 1959; Kasha and Pullman, 1962). This molecule may actually have been rather small, and it is remarkable that the mechanisms of replication and mutation may be sufficient to explain the synthesis of the enormously long DNA molecules which are characteristic for the higher organisms. If the small polynucleotide has the power of self-replication, it probably also has under certain conditions the ability of doubling its own length by adding the copy to the original. The process could be repeated and would lead to a long DNA molecule containing the same genetic message associated with the original polynucleotide repeated over and over again. However, due to influence of proton transfer, the base sequence may be altered with time and get a more complicated character. In the struggle for the nutrition material, the principle of "natural selection" would then start working.

It would take us too far to go any deeper into the theory of evolution here, and we will only give a few references. It is evident, however, that the question of the base composition of DNA and the order of the base sequence represent very important problems in this connection. The question of sequence distribution and neighbor effects has been treated by Simha and Zimmerman (1962), and the problem as to the evolution of base composition has been studied by Freese (1959a, b). In a study of the process of evolution, B. and A. Pullman (1962b) have carried out a quantum-mechanical investigation of the four bases and have found that cytosine has very likely the highest probability to go over into its tautomeric form. The tentative studies of proton tunneling in DNA carried out so far indicate that the G—C pair seem to be considerably less than the A—T pair with respect to proton tunneling, and this phenomenon could then explain the well-known fact that DNA molecules of the higher organisms have a larger content of A—T pairs than G—C pairs.

VIII. Problem of Aging

Biological organisms consisting of a single cell do not "age." They propagate by cell division and, since one cell suddenly becomes two, the two become four, etc., a cell lacks individual identity and cannot age in the ordinary sense. If such a cell becomes damaged, it will either die directly or lose out in the struggle for existence in competition with normal cells. This implies that a system of such cells will usually not show any loss of vigor with time.

The same applies to certain cells taken from multicellular organisms. In his classical experiments, Alexis Carrel kept a culture of chicken embryo heart cells for 34 years and at the end the cells showed no sign of aging. Many authors have interpreted these experiments so that the phenomenon of biological aging cannot depend on the cells but is instead essentially confined to the "structural matrix" of the organism which forms the framework for the cells.

Aging is hence a phenomenon which is characteristic for the multicellular organisms, and it seems to have to do with the organism itself, i.e., the superstructure of cells which constitute a biological individual. In discussing the problem of the cause of aging, one has often pointed to a series of effects common for large groups of living organisms, e.g., arteriosclerosis in mammals. Even if most of these effects seem to depend rather strongly on time, it is clear that the organism is originally able to handle these complications, since otherwise the accumulation of these effects would probably be lethal much earlier. Perhaps aging can then best be described as an expression for the organisms increasing loss of control of the entire metabolism, and the phenomenon should then be traced back to the inability to produce the enzymes necessary to regulate the normal processes. It seems hence very likely that aging may be due primarily to a change or loss of the genetic information.

Such a theory of aging was proposed in 1959 by Szilard (1959, 1960), who assumed that the genes are subject to random "aging hits" which make them inactive with respect to the synthesis of a specific protein having a certain catalytic activity. By assuming that the probability that a chromosome of a cell suffers an "aging hit" per unit time is a constant which is the same for all chromosomes, Szilard could make an interesting study of the ratio of boys and girls, at birth, as a function of the age of the father.

Let us now study the problem of aging from the point of view that the genetic code is carried by the DNA molecule. The material in the various parts of the body of an organism is constantly renewed at a certain rate characteristic for the species and the types of cells involved, which means that the somatic cells are undergoing repeated cell replications (mitosis). The analysis of the genetic code in the Watson–Crick model has shown that, depending on the "tunnel effect," a "pure" genetic message does not exist, and that there is always a finite probability error built into even the normal state of the DNA molecule. In addition, there are incorporation errors which may occur at each replication, and "transcription errors" which may be induced during the synthesis of messenger RNA in the deep groove of DNA. Since all these types of errors may cause a certain amount of loss of genetic information in each DNA replication, it seems hence very likely that the phenomenon of aging would at least partly depend on the accumulation of such loss of genetic information.

As emphasized in the introduction, the differentiation of the cells may play a fundamental role in connection with the phenomenon of aging. In tissues where the cells divide frequently, a somatic mutation will usually cause little harm, since the corresponding cell will immediately lose out in competition with the normal cells. On the other hand, all multicellular organisms seem to have certain tissues in which the cells have stopped their replication at a certain stage of life, for instance, nerve cells and brain cells, and it seems as if the phenomenon of aging to a large extent would be concentrated in such cells (Curtis, 1963).

In the absence of replication errors, the loss of the genetic code should then depend essentially only on tunneling errors and on transcription errors related to the synthesis of messenger RNA. In addition to the loss of the genetic code, there may also be other errors accumulating in various parts of the chain regulating the protein synthesis (Bjorksten, 1958, 1962, 1963; Bjorksten *et al.*, 1960, 1962; Orgel, 1963).

It seems to be significant that both the spontaneous errors in DNA and the transcription errors are ultimately dependent on proton transfer through the tunnel effect. From this point of view, aging is then analogous to a radioactive decay process with a rather long half-life, with the important difference that proton transfer in DNA is highly temperature-dependent and influenced by the electrical environment, etc. Perhaps it is not a mere coincidence that one also measures the age of archeological pieces and the earth itself by means of the same type of tunnel effect through the radioactivity. It is one of those rare phenomena which span all orders of magnitude with respect to time. It should finally be mentioned that the disintegration of dead matter and particularly solid state through corrosion is also believed to depend on a tunneling phenomenon.

IX. Some Aspects on the Connection between the DNA Structure and the Development of Tumors and Cancer

A. Spontaneous Tumors

The growth of an individual is a highly refined balance between factors which enhance the cell duplication and other factors which limit this duplication so that the organism takes a specific shape. The entire process is stimulated and controlled by various enzymes, hormones, and other control substances (Szent-Györgyi *et al.*, 1963), and there is a feedback from the environment about which presently we know very little. If there is a *somatic mutation*, i.e., a change of the genetic code in a DNA molecule in the body of an organism, the change may influence the protein synthesis and the balance between the enhancing and controlling enzyme actions in the growth cycle (Levan, 1959; Arley, 1960; Arley *et al.*, 1952, 1962; Iversen and Arley, 1953). Actually, the new genetic code may lead to the development of a "new individual" within the individual, i.e., a tumor. Fortunately not all types of tumors are malignant. For a survey of the problem we refer to the proceedings of the Ciba Foundation Symposium, 1958, "Carcinogenesis, Mechanisms of Action" (Wolstenholme and O'Connor, 1959) and to the proceedings from the 8th International Cancer Congress in 1962 (Maisin, 1963).

Cancer will here be described as the growth of such abnormal cells in the living organism which have a higher rate of metabolism than the normal cells and which hence may take over the normal material in large areas of the organism and form malignant tumors. It is characteristic for the cancer cells that they often seem to lose the specificity of the tissue from which they originate. In the *deletion hypothesis,* cancer is essentially assumed to be caused by the fact that the growth-controlling factors—enzymes, hormones, etc.—for some reason have been deleted. This may depend on an error in the DNA molecule itself, e.g., through a somatic mutation, but it may also depend on other errors along the assembly lines for the growth-controlling molecules of perhaps more temporary nature.

In this connection, it is important to understand how the cell differentiation in a multicellular organism is achieved. One knows from experience that the genetic information contained in the system of DNA molecules in a cell nucleus is complete, but also that only a small part is utilized for protein synthesis in a differentiated cell—only such structural proteins and enzymes as are needed are produced. This means that the genetic information in a cell nucleus may be compared with a book with many chapters, in which only a specific chapter is "read" and utilized for controlling the metabolism in a particular type of cell (Platt, 1962). All other chapters are "closed" which is dependent on the fact that the deep groove in DNA may be blocked by certain molecules. For a review of the details of the procedure regulating the reading of the genetic code, we refer to some recent papers (Jacob and Monod, 1961; Rich, 1962; Platt, 1962).

One may wonder how a cell obtains such a specific message as "read chapter one," "read chapter two," etc. The simplest mechanism is based on the idea that a certain information I, which originally may have been obtained, for example, by reading a single strand in DNA, is not duplicated before a cell division but is instead divided into two parts, $I = I_1 + I_2$, going into different daughter cells. This leads to a replication scheme of the type illustrated in Fig. 77. When the material in a tissue is renewed, it is essential that both the DNA molecule as well as the molecules regulating the reading of DNA are properly duplicated before the cell division. If some of the molecules blocking the reading of the deep groove of DNA would be deleted, the cell would lose part of its differentiation and get a foreign appearance in the tissue. However, since the cell does not contain any molecules which are foreign to the organism, the usual immunological reactions are not immediately induced. The occurrence of tumors and cancer tells us that certain dedifferentiated cells have a larger replication power than the normal differentiated cells, and there are indications that the growth power increases when the molecules or control factors blocking the deep groove of DNA are removed. From this point of view, a DNA molecule by itself without its proper framework would be very dangerous for the organism.

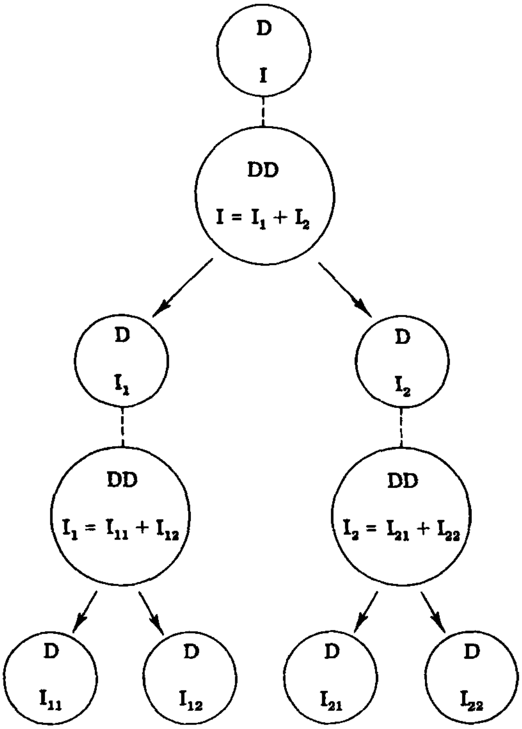


Fig. 77. Cell replication leading to a differentiation of the cells. D = DNA, I = Information for reading DNA.

It has been discussed whether cancer is a virus disease, since cancer may in certain cases be induced by injection of highly filtered extract from a cancer tumor on a healthy individual of the same species. Today, one can understand these experiments in terms of the "transformation principle" discussed in Section II, and it is clear that the extract may contain DNA molecules distorted in a carcinogenic way. Today, it seems rather unlikely that the main forms of cancer are originally caused by some external virus, i.e., by external RNA or DNA molecules in protein overcoats existing in free form. It seems much more probable that the special "cancer virus" is of intracellular nature and consists of the organisms's own RNA or DNA molecules without the proper regulators or with a changed genetic message which enables them to change the protein synthesis in their favor and to take over the metabolism and cause cancer. This view explains also why the cancer pattern can be so easily transported within the organism and give rise to daughter tumors (metastasis).

It should be emphasized that this picture does not prevent the existence of a real "virus cancer" caused by the fact that an external virus can come into the cell and influence the original DNA molecule giving rise to a somatic mutation of a rather large order, i.e. acting essentially as a mutagen (Blum, 1963), or attack other parts of the assembly lines for the molecules regulating the replication and reading of DNA.

Basing the discussion on the hypothesis that all the chemical processes in a cell including the replication and transcription of DNA are ultimately controlled by DNA itself, one is lead to the conclusion that *somatic mutations* may represent at least one of the main causes for the spontaneous occurrence of tumors and cancer. Such somatic mutations may be established through incorporation errors, transcription errors, or double proton tunneling in the base pairs of DNA and, in all these cases, the errors are associated with proton transfer over a distance of about $1\ \text{Å} = 10^{-8}$ cm which is facilitated by the quantum-mechanical tunnel effect. Since the causes for somatic mutations leading to the formation of tumors are essentially the same as in the aging process, there ought to be a clear correlation between age and the occurrence of tumors, with the important difference that tumors occur mainly in tissues which renew themselves by cell duplication, whereas aging is usually limited to cells which, after a certain period in life, do not replicate at all.

B. Induced Tumors

In the treatment of the problem of carcinogenesis, Berenblum (1959) and others have developed a theory based on the existence of two stages:

(1) initiation;
(2) promotion.

The initiation is usually assumed to consist of a somatic mutation, i.e., a change of the genetic code in DNA, induced by chemicals, by radiation, or by other means. It should be observed that, even if the new code does not contain a highly carcinogenic message, one may further disturb the balance between the growth-enhancing and the growth-controlling compounds regulating the replication and reading of DNA by means of suitable chemicals during the promotion stage. This may start the growth of a malignant tumor which may then develop quickly through a chain reaction. The treatment of the promotion is essentially outside the framework of this paper, and we will confine ourselves to consider only the somatic mutations.

Somatic mutations may be induced by essentially the same mechanisms as ordinary mutations, i.e., by chemicals, by radiation, and by certain other physical means. The chemicals should have the property of being transferred to the region of the DNA replication, and one would hence expect that a special group is formed by those which may be easily attached to the proteins which catalyze this replication (cf. also Mason, 1960).

An extensive theoretical study of the carcinogenic activity of the catacondensed hydrocarbons has been carried out by the groups in Paris. By investigating the connection between charge and bond orders and carcinogenic activity, the Pullmans (A. and B. Pullman, 1955a, b, 1959, 1961a, b, 1962c; B. and A. Pullman, 1963) have found that there are two regions of special interest in these hydrocarbons denoted as K and L regions, illustrated in Fig. 78, and that the molecule is carcinogenic if the chemical activity in the K region is sufficiently high and the activity in the L region is sufficiently low (cf. Coulson, 1953). Experiments by Heidelberger (1959) and by Mme. Daudel *et al.* (1960a, b, c, 1961, 1962, 1963; Daudel and Ferraz, 1963; Daudel and Prodi, 1963; Jacquier and Daudel, 1964) have indicated that these conditions may be connected with the ability of the molecules to get attached to certain proteins. Since molecules of this type may enter electronic donor-acceptor reactions with the nucleotide bases, one can expect that they may induce mutations if they reach the immediate neighborhood of the DNA molecule. Little is known about the actual mechanism, however—whether it is based on "sandwiching" of the mutagen between base pairs or not (B. Pullman, 1962; Lerman, 1963; Liquori, 1963; Boyland, 1964; Heidelberger, 1963)—and it should be

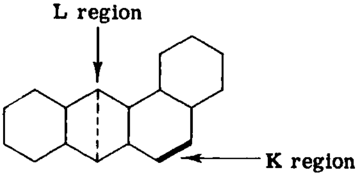


Fig. 78. Aromatic hydrocarbon showing typical K and L regions according to A. and B. Pullman.

observed that this theoretical study was started in Paris before the important role of the nucleic acids in genetics was realized. The information was gained solely by studying the *correlation* between electronic structure and carcinogenic activity, and the results achieved in this way are remarkable.

The inducing effect of radiation on somatic mutations seems to depend on ionization, essentially in the same way as in the ordinary mutations. Again the genetic code is disturbed by proton transfer and, under certain circumstances, a carcinogenic message may result.

It is interesting to observe that, if the electrostatic picture of the hydrogen bonding is true, one can expect that an outer electrostatic potential may induce proton tunneling and disturb the genetic message. This is a clear *physical effect* without any chemical interaction, and it could be caused, for instance, by means of a dipole layer introduced for a sufficiently long time in a tissue. Since plastic films are often statically charged—they are technically used as "electrets"—and metals in liquids assume electrochemical potentials, one could expect to obtain results with rather small means. However, it is too early to say whether there is a definite correlation between these electrostatic effects and Oppenheimer's experimental results (see, e.g., Alexander and Hornig, 1959) concerning the high carcinogenic activity of plastic films and certain metals embedded in small sheets in tissues, but the connection seems certainly interesting and worthwhile investigating. It should be observed that no carcinogenic effect was obtained if the plastic material was introduced as a powder, so it seems as if the effect would have a physical origin.

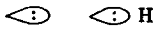
The medical treatment of cancer has to depend on the difference between cancer cells and normal cells. It has been found that the former are more sensitive to radiation than the latter, and this may depend on the fact that radiation causes additional errors in the base sequences of the DNA molecules involved, and that a too large deviation from the normal sequence may lead to DNA molecules which are not reproducible by a growth cycle. However, the radiation influences the growth cycle and the metabolism in many other ways, and it would be interesting to study the absorption spectrum of cancer cells or nuclei in comparison to normal cells or nuclei to see whether one could get more selective effects by limiting the radiation to specific frequencies or by making it "monochromatic."

The chemotherapy of cancer has to be directed towards stopping the growth cycle of the cancer cells leaving the normal cells as little damaged as possible. Again the base sequences of the DNA and RNA molecules are probably the weakest points of the cancer cells, and by introducing "base analogs," i.e., bases similar to the normal bases but without the proper action in the protein synthesis, as nutrition material in the cancer cells, one can hope to stop their duplication. The base analogs should be such that they are accepted as normal building material in the DNA duplication, but they should then be inactive in the formation of RNA leading to errors deactivating the protein synthesis, or they should show a tendency to destroy the proton code (see Section VII). The difficulties are here connected with the fact that many "base analogs" are also highly poisonous to the normal cells. Even other points in the cancer metabolism may be attacked by suitable chemicals, e.g., the enzyme actions, and the possibility of disturbing the action of DNA polymerase in the cancer cells should particularly be investigated. In connection with induced cancer, the stage of promotion should be given particular attention.

Other differences between the cancer cells and the normal cells, e.g., a difference in permeability of their membranes, may be used to prepare the cancer cells for the final attack by radiation, base analogs, or other chemicals. The difficulty in all medical treatments of cancer lies in the fact that, if a single cancer cell or erroneous DNA molecule is left, the cancer pattern may develop again as it started.

Our study makes it to a certain extent rather likely that the ultimate cause of spontaneous cancer—the "principle of cancer"—may be a submolecular phenomenon associated with protons changing their places over a distance of the order of magnitude of 10^{-8} cm through potential barriers along hydrogen bonds by means of the tunnel effect either in the base pairs themselves or in connection with incorporation and transcription errors. However, it should be emphasized that such a model of the cause of cancer is highly oversimplified. The DNA molecule may contain the essential genetic information, but RNA molecules and proteins are also inherited by the daughter cells after a cell duplication. The protein synthesis and its feedback on the cell duplication is a highly complicated biochemical process of which one has only a limited knowledge, and the role of the deep groove in DNA in the replication and transcription mechanisms and their regulation is still a deep secret, even if one can anticipate that many points may be relevant also in connection with cancer.

In the DNA and RNA molecules, nature has found an excellent mechanism for storing long-time information in the form of a proton-electron pair code of the type



Such a storage is highly similar to the storage mechanism in a modern electronic computer, with the difference that the "microcomponents" used by nature can hardly be beaten by any technical development [cf. the criticism by Commoner (1962).] The question is whether nature is using the same type of mechanism also for storing short-time information of the type usually referred to as "memory," consisting of the ordinary conscious memory, the reflex memory, and the "chemical memory" rendering immunity against certain diseases. There are indications that nature is utilizing DNA and RNA molecules for this purpose also, but where in the cell they are situated or how they are organized is still not known. Of course, molecules of this type may also be of importance in regulating the cell processes, and it seems likely that ultimately the control mechanism depends on a complicated interplay between nucleic acids and proteins in various parts of a single cell and in different cells of the same organism.

X. Discussion

A. Summary

Let us now summarize the main points discussed in this paper. Deoxyribonucleic acid (DNA) is considered as the hereditary substance and, according to Watson–Crick's model, the genetic message is contained in a proton-electron pair code which is situated well-hidden and shielded in the middle of a double helix. The code consists actually of two complementary pieces of "lock and key" type which together have a great deal of stability. In the replication process, the code is opened only momentarily to find the correct partners for the doubling of the genetic message. All these precautions give the genetic code an unusual stability and explain its ability to preserve a genetic message intact over thousands of years. The genetic information is transferred to the cell by means of the formation of messenger RNA in the deep groove in DNA, but it is not yet known whether, during the transcription procedure, the genetic code is momentarily opened up or the message is read in an extra "copy" associated with the 6 positions of the base pairs.

In this paper we have pointed out that, since the protons are not classical particles but "wave packets" obeying the laws of modern quantum theory, the genetic code cannot—in spite of all precautions—be 100% stable. Due to the quantum-mechanical "tunnel effect," there is always a small but finite probability that the protons in a base pair will change place, alter the genetic code, and give rise to a mutation. Similarly, proton transfer facilitated by the tunnel effect may occur in connection with the replication and transcription procedures. This implies that proton jumps over a distance of $1 \text{ Å} = 10^{-8}$ cm may be one of the driving forces in the evolution of living organisms on the earth. In a multicellular organism, the same "tunnel effect" may lead to a loss of genetic information which may be the primary cause of the phenomenon of aging. Since the proton tunneling and the incorporation of double errors further lead to somatic mutations, the phenomenon may also be at least partly responsible for the occurrence of spontaneous tumors and cancer. The proton transfer is of particular importance in understanding induced phenomena. It is evident that a model of these biological phenomena where all the emphasis is put on the DNA molecule must be oversimplified since there are certainly also other cell constituents which play an important role in these connections. We believe, however, that the picture serves a meaningful purpose as a first approximation.

Hydrogen bonds and proton transfers are of importance not only in the nucleic acids but also in the proteins and their activity, and hence they play a fundamental role in biochemistry (Löwdin, 1964b). It has been emphasized that the proton transfer is greatly facilitated by the "tunnel effect" and that, even if the quantum theory of this effect is comparatively simple, the proper evaluation of tunneling times requires highly accurate knowledge of the electronic and protonic structure of the compounds involved. This depends on the fact that the tunneling probability per second c for a proton is highly sensitive to changes in the tunneling distance a_0 and the tunneling barrier V_0. In biochemistry, it may be convenient to express a_0 in angstroms ($= 10^{-8}$ cm) and V_0 in electron volts, and Table II shows that, if the parameter $a_0\sqrt{V_0}$ is shifted from 0.1 to 2.0, the tunneling time τ spans almost over our entire time scale from very fast ($10^{-11.5}$ sec) to very slow (10^{17} sec), the latter figure corresponding to the present estimate of the age of the universe. Only certain nuclear events may be faster (10^{-23} sec). The escape of a proton through a potential barrier is further illustrated in Fig. 79 as a function of a logarithmic time scale indicating also the time period for some fundamental events. One sees clearly how a small shift in the parameter $a_0\sqrt{V_0}$ is going to influence greatly the entire tunnel phenomenon. This implies also that a proper treatment of this effect is going to require essential refinements of many of the methods presently used in quantum chemistry.

B. Molecular Biology and Quantum Biology

A few words should be added about the implications of the new approach of discussing biological phenomena on a molecular and submolecular level. It is clear that there is a considerable "language difficulty" due to the fact that the terminology and conceptual structure developed in one field, say on the level of cellular organization, may not at all be convenient for the description in another field, say on the molecular level. Part of the importance of the Watson–Crick model lies in the fact that it tries to bridge the gap between the life sciences and molecular theory.

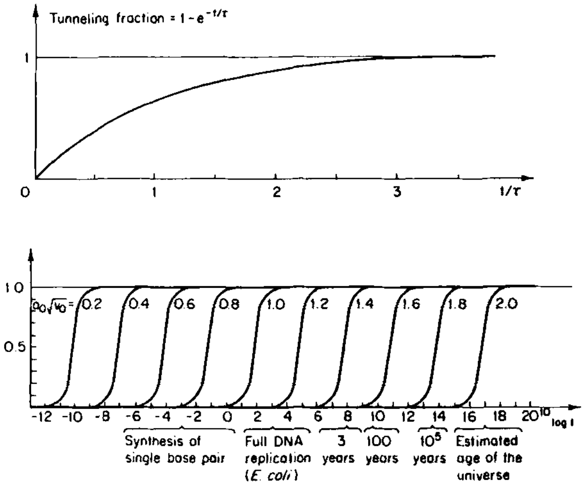


Fig. 79. Fraction of a proton escaping through a potential barrier with tunneling distance a_0 Å and height V_0 eV above the tunneling level.

During the last three decades, there has been a very strong tendency towards the unification of all the natural sciences. The development has also shown that many important achievements in medicine depend on the basic research carried out in the new field of molecular biology. From the point of view of health, the living systems are essentially huge chemical systems, which are in their normal or abnormal behavior subject to the laws of physics and chemistry. In the recent development of the biosciences, more and more emphasis has been put on the fundamental molecular structure of the living systems and on the detailed electronic and protonic structure of the molecules involved. The Pullmans (A. and B. Pullman, 1960, 1962a, b, c; B. and A. Pullman, 1963) have pointed out that the flexibility and the extremely high mobility of the living systems seem in many cases to be connected with the properties of the "mobile electrons" of the conjugated systems, which occur as essential constituents in the giant molecules in the living systems.

The electrons and protons are fundamental particles which do not obey the laws of classical physics but the laws of modern quantum mechanics.

The electronic and protonic structure of biologically interesting molecules has hence to be treated by quantum chemistry, and this seems to lead to the opening of a new field which should perhaps be called "quantum biology." The principles of quantum mechanics are of fundamental importance for treating not only the ground state and the excited states of conjugated systems and polynucleotides, but also the biologically interesting molecules in general, the problem of energy storage and energy transfer, and many other basic problems in biophysics and biochemistry which are now treated under the common name of "submolecular biology" (Szent-Györgi, 1960; Kasha and Pullman, 1962; Setlow and Pollard, 1962). The modern theory of the structure of matter has through quantum mechanics rendered a unification of the fundaments of physics and chemistry which was previously inconceivable, and it seems as if the biosciences would be next in turn to join the same basis.

ACKNOWLEDGMENTS

The author would like to thank Dr. Olle Mårtensson for fruitful discussions and for most valuable help in going through the literature and collecting the references, and Fil. lic. Klaus Appel and Fil. kand. Karl-Fredrik Berggren for skillful assistance in preparing the figures and the tables. He is further indebted to Dr. J. Ladik for his excellent series of lectures on the Watson-Crick model of DNA and its biological functioning in Uppsala in the spring of 1962, which has been an important source of inspiration. Important information concerning the biochemical aspects has been obtained from Dr. Hans G. Boman, concerning the tautomeric equilibria from Dr. Göran Bergson, and concerning the double-well potential in DNA from Dr. Robert Rein and Prof. Frank Harris, which is hereby gratefully acknowledged.

Special thanks should finally be expressed to Prof. B. Pullman and Dr. A. Pullman for many valuable and stimulating discussions on these problems during the Summer Institute in Quantum Chemistry and Solid-State Physics in Uppsala in 1962.

REFERENCES

ALEXANDER, P., and HORNIG, E. S. (1959). *Ciba Found. Symp., Carcinogenesis, Mechanisms Action*, p. 12.

ALI, M. A., and WOOD, R. F. (1962). *Uppsala Quantum Chem. Group Tech. Note* No. 75.

ARLEY, N. (1960) *Proc 4th Symp. Math. Statist. Probability, Berkeley, 1960*, p. 18.

ARLEY, N., and IVERSEN, S. (1952). *Nature* **169**, 410.

ARLEY, N., and SETÄLÄ, K. (1962). *Nature* **194**, 436.

AVERY, O. T., MACLEOD, C. M., and MCCARTY, M. (1944). *J. Exptl. Med.* **79**, 137.

BELEZNAY, F., BICZÓ, G., and LADIK, J. (1964). *Phys. Letters* **11**, 234.

BELL, C. L., and BARROW, G. M. (1959). *J. Chem. Phys.* **31**, 300.

BELOZERSKY, A. N., and SPIRIN, A. S. (1958). *Nature* **182**, 111.

BENZER, S., and CHAMPE, S. P. (1962). *Proc. Natl. Acad. Sci. U.S.* **48**, 1114.

BERENBLUM, I. (1959). *Ciba Found Symp., Carcinogenesis, Mechanisms Action*, p. 55.

BERG, P. (1961). *Ann. Rev. Biochem.* **30**, 293.

BICZÓ, G., LADIK, J., and GERGELY, I. (1964). *Phys. Letters* **13**, 317.

BJORKSTEN, J. (1958). *J. Am. Geriat. Soc.* **6**, 740.

BJORKSTEN, J. (1962). *J. Am. Geriat. Soc.* **10**, 125.

BJORKSTEN, J. (1963). *Gerontologia* **8**, 179.

BJORKSTEN, J., and ANDREWS, F. A. (1960). *J. Am. Geriat. Soc.* **8**, 632.

BJORKSTEN, J., ANDREWS, F. A., BAILEY, J., and TRENK, B. (1960). *J. Am. Geriat. Soc.* **8**, 37.

BJORKSTEN, J., ANDREWS, F. A., PRAHL, H. F. (1962). *Finska Kemistsamfundets Medd.* **71**, 1.

BLOCH, D. P. (1955). *Proc. Natl. Acad. Sci. U.S.* **41**, 1058.

BLUM, H. F. (1963). *Nature* **199**, 155.

BOYLAND, E. (1964). *In* "Electronic Aspects of Biochemistry" (B. Pullman, ed.), p. 155. Academic Press, New York.

BOZÓKY, L., KISZELY, GY., HOFFMANN, T. A., and LADIK, J. (1963). *Nature* **199**, 1306.

BRACHET, J. (1950). "Chemical Embryology." Wiley (Interscience), New York.

BRACHET, J. (1960). *Nature* **186**, 194.

BRENNER, S. (1957). *Proc. Natl. Acad. Sci. U.S.* **43**, 687.

BRENNER, S., BARNETT, L., CRICK, F. H. C., and ORGEL, A. (1961). *J. Mol. Biol.* **3**, 121.

BRILLOUIN, L. (1926a). *Compt. Rend.* **183**, 24.

BRILLOUIN, L. (1926b). *J. Phys. (Paris)* **7**, 353.

BURMA, D. P. (1962). *J. Sci. Ind. Res. (India)* **21A**, 549.

BURNET, F. M., and STANLEY, W. M., eds. (1959). "The Viruses," 3 vols. Academic Press, New York.

BUTLER, J. A. V. (1963). *Nature* **199**, 68.

CAIRNS, J. (1963). *J. Mol. Biol.* **6**, 208.

CASPERSSON, T. (1942). *Naturwiss.* **29**, 33.

CASPERSON, T. (1950). "Cell Growth and Cell Function," Norton, New York.

CAVALIERE, L. F., and ROSENBERG, B. H. (1961). *Biophys. J.* **1**, 317, 323, and 337.

CHANTRENNE, H. (1961). "Biosynthesis of Proteins." Pergamon Press, Oxford.

CHAPEVILLE, F., LIPMANN, F., VON EHRENSTEIN, G., WEISBLUM, B., RAY, W. J., JR., and BENZER, S. (1962). *Proc. Natl. Acad. Sci. U.S.* **48**, 1086.

CHARGAFF, E. (1950). *Experientia* **6**, 201.

CHARGAFF, E. (1955). *In* "The Nucleic Acids" (E. Chargaff and J. N. Davidson, eds.), Vol. I, p. 307. Academic Press, New York.

CHARGAFF, E. (1957). *In* "The Chemical Basis of Heredity," Symposium (W. D. McElroy and B. Glass, eds.), p. 521. John Hopkins Press, Baltimore, Maryland.

CHARGAFF, E. (1963). *In* "Essays on Nucleic Acid" (W. D. McElroy and B. Glass, eds.), Johns Hopkins Press, Baltimore, Maryland.

CHARGAFF, E., MAGASANIK, F., VISCHER, E., GREEN, S., DONIGER, R., and ELSON, D. (1950). *J. Biol. Chem.* **186**, 51.

COMMONER, B. (1962). *In* "Horizons in Biochemistry" (M. Kasha and B. Pullman, eds.), p. 319. Academic Press, New York.

CORRENS, C. (1900). *Ber. Deut. Botan. Ges.*

COULSON, C. A. (1953). *Advan. Cancer Res.* **1**, 1.

COULSON, C. A. (1959). *Proc. Intern. Symp. Hydrogen Bonding, Ljubljana 1957*, p. 339. Pergamon Press, Oxford.

COULSON, C. A., and LONGUET-HIGGINS, H. C. (1947a). *Proc. Roy. Soc.* **A191**, 39.

Coulson, C. A., and Longuet-Higgins, H. C. (1947b). *Proc. Roy. Soc.* **A192**, 16.

Coulson, C. A., and Longuet-Higgins, H. C. (1948a). *Proc. Roy. Soc.* **A193**, 447 and 456.

Coulson, C. A., and Longuet-Higgins, H. C. (1948b). *Proc. Roy. Soc.* **A195**, 188.

Crick, F. H. C. (1957). *In* "The Chemical Basis of Heredity," Symposium (W. D. McElroy and B. Glass, eds.), p. 532. Johns Hopkins Press, Baltimore, Maryland.

Crick, F. H. C. (1958). *Symp. Soc. Exptl. Biol.* **12**, 138.

Crick, F. H. C. (1962). Uppsala Lectures, unpublished.

Crick, F. H. C., and Kendrew, J. C. (1957). *Advan. Protein Chem.* **12**, 133.

Crick, F. H. C., and Watson, J. D. (1954). *Proc. Roy. Soc.* **A223**, 80.

Crick, F. H. C., Griffith, J. S., and Orgel, L. E. (1957). *Proc. Natl. Acad. Sci. U.S.* **43**, 416.

Crick, F. H. C., Barnett, L., Brenner, S., and Watts-Tobin, R. J. (1961). *Nature* **192**, 1227.

Curtis, H. J. (1963). *Science* **141**, 686.

Darwin, C. R. (1859). "Origin of Species."

Daudel, P., and Ferraz, A. M. (1963). *Compt. Rend.* **256**, 538.

Daudel, P., and Prodi, G. (1963). *Acta Unio Intern. Contra Cancrum* **19**, 522.

Daudel, P., Prodi, G., and Chenon, B. (1959a). *Compt. Rend.* **248**, 3238.

Daudel, P., Vallée, G., and Vasquez, R. (1959b). *Compt. Rend.* **248**, 1880.

Daudel, P., Chenon, B., Buu-Hoi, N. P., Jacquignon, P., Lacassagne, A., Prodi, G., Vallée, G., Vasquez, R., and Zajdela, F. (1960a). *Bull. Soc. Chim. France* **42**, 135.

Daudel, R., Lefebvre, R., and Moser, C. (1960b). "Quantum Chemistry." Wiley (Interscience), New York.

Daudel, P., Nectoux, F., Pichat, L., and Prodi, G. (1960c). *Compt. Rend.* **251**, 1049.

Daudel, P., Prodi, G., Fabel, A., and Morniche, M. C. (1961). *Compt. Rend.* **253**, 593.

Daudel, P., Muel, B., Lacroix, G., and Prodi, G. (1962). *J. Chim. Phys. (Paris)* p. 263.

Daudel, P., Jacquier, A., Nakajima, A., and Fossar, N. (1963). *Compt. Rend.* **257**, 2910.

Davidson, J. N. (1960). "The Biochemistry of the Nucleic Acids," 4th ed. Methuen, London (see particularly pp. 41–44).

Davies, M. (1959). *Proc. Intern. Symp. Hydrogen Bonding, Ljubljana 1957*, p. 393. Pergamon Press, Oxford.

Delbrück, M. (1955). *Proc. Natl. Acad. Sci. U.S.* **40**, 783.

Delbrück, M., and Stent, G. S. (1957). *In* "The Chemical Basis of Heredity," Symposium (W. D. McElroy and B. Glass, eds.), p. 699. Johns Hopkins Press, Baltimore, Maryland.

Delbrück, M., Timoféëff, N. W., and Zimmer, K. G. (1935). *Nachr. Biol. Ges. Wiss. Göttingen* **1**, 189.

de Vries, H. (1900). *Ber. Deut. Botan. Ges.*

Doty, P. (1959). *Rev. Mod. Phys.* **31**, 107.

Dounce, A. L. (1952). *Enzymologia* **15**, 251.

Dounce, A. L. (1962). *J. Theoret. Biol.* **2**, 152.

Eck, R. V. (1962). *J. Theoret. Biol.* **2**, 139.

Eck, R. V. (1963). *Science* **140**, 477.

Esaki, L. (1960). *In* "Solid State Physics in Electronics and Telecommunications" (M. Désirant and J. L. Michiels, eds.), Vol. 1, p. 514. Academic Press, New York.

Fischer-Hjalmars, I., and Grahn, R. (1958). *Acta Chem. Scand.* **12**, 584.

Fong, P. (1964). *Proc. Natl. Acad. Sci. U.S.* **52**, 641.

Fowler, R. H., and Nordheim, L. (1928). *Proc. Roy. Soc.* **A119**, 173.

Fraenkel-Conrat, H. (1959). *Harvey Lectures* **53**, 56.

Fraenkel-Conrat, H., Singer, B. A., and Williams, R. C. (1957). *In* "The Chemical Basis of Heredity," Symposium (W. D. McElroy and B. Glass, eds.), p. 501. Johns Hopkins Press, Baltimore, Maryland.

Fraenkel-Conrat, H., Tsugita, A., Nirenberg, M .W., and Matthaei, J. H. (1962). *Proc. Natl. Acad. Sci. U.S.* **48**, 846.

Frank, N. H., and Young, L. A. (1931). *Phys. Rev.* **38**, 80.

Freese, E. (1959a). *J. Mol. Biol.* **1**, 87.

Freese, E. (1959b). *Proc. Natl. Acad. Sci. U.S.* **45**, 622.

Freese, E. (1962). *J. Theoret. Biol.* **3**, 82.

Freudenthal, H. (1958). *Koninkl. Ned. Akad. Wetenschap., Proc.* **A61**, 253.

Fruton, J. S., and Simmonds, S. (1961). "General Biochemistry." Wiley, New York.

Furberg, S. (1950). *Acta Cryst.* **3**, 525.

Furberg, S. (1952). *Acta Chem. Scand.* **6**, 634.

Gamow, G. (1928a). *Z. Physik* **51**, 204.

Gamow, G. (1928b). *Z. Physik* **52**, 510.

Gamow, G. (1954a). *Nature*, **173**, 318.

Gamow, G. (1954b). *Kgl. Danske Videnskab Selskab, Biol. Medd.* **22**, 3.

Gamow, G., and Ycas, M. (1955). *Proc. Natl. Acad. Sci. U.S.* **41**, 1011.

Gamow, G., Rich, A., and Ycas, M. (1956). *Advan. Biol. Med. Phys.* **4**, 23.

Gavrilov, V. Y., and Zograf, V. N. (1963). *Soviet Phys.—Usp.* **5**, 634.

Geiduschek, E. P., Moohr, J. W., and Weiss, S. B. (1962). *Proc. Natl. Acad. Sci. U.S* **48**, 1078.

Golomb, S. W., Welch, L. R., and Delbrück, M. (1958). *Kgl. Danske Videnskab. Selskab, Biol. Medd.* **23**, 9.

Grahn, R. (1959). *Arkiv Fysik* **15**, 257.

Grahn, R. (1961). *Arkiv Fysik* **19**, 147.

Grahn, R. (1962). *Arkiv Fysik* **21**, 1, 13, and 81.

Griffith, F. (1928). *J. Hyg.* **23**, 113.

Guild, W. R., and Robinson, M. (1963). *Proc. Natl. Acad. Sci. U.S.* **50**, 106.

Gurney, R. W., and Condon, E. U. (1929). *Phys. Rev.* **33**, 127.

Haas, C., and Hornig, D. F. (1959). *J. Chem. Phys.* **32**, 1763.

Hadzi, D. (1959). *Proc. Intern. Symp. Hydrogen Bonding, Ljubljana 1957.* Pergamon Press, Oxford.

Hall, B. D., and Spiegelman, S. (1961). *Proc. Natl. Acad. Sci. U.S.* **47**, 137.

Haselkorn, R., Fried, V. A., and Dahlberg, J. E. (1963). *Proc. Natl. Acad. Sci. U.S* **49**, 511.

Hayashi, M., Hayashi, M. N., and Spiegelman, S. (1963). *Science* **140**, 1313.

Heidelberger, C. (1959). *Ciba Found Symp., Carcinogenesis, Mechanisms Action*, p. 179.

Heidelberger, C. (1963). Unpublished.

Heitler, W., and London, F. (1927). *Z. Physik* **44**, 455.

Hershey, A., and Chase, M. (1952). *J. Gen. Physiol.* **36**, 39.

Herzberg, G. (1950). "Spectra of Diatomic Molecules," 2nd ed. Van Nostrand, Princeton, New Jersey (see particularly p. 409).

Hirschfelder, J. O., and Wigner, E. (1939). *J. Chem. Phys.* **7**, 616.

Hoagland, M. B. (1959). *Brookhaven Symp. Biol.* **12**, 40.

HOAGLAND, M. B., ZAMECNIK, P. C., and STAPHENSEN, M. L. (1959). *In* "A Symposium on Molecular Biology" (R. E. Zirkle, ed.), p. 105. Univ. of Chicago Press, Chicago, Illinois.

HOFACKER, L. (1958). *Z. Naturforsch.* **13a**, 1044.

HOFACKER, L. (1959). *Proc. Intern. Symp. Hydrogen Bonding, Ljubljana 1957*, p. 375. Pergamon Press, Oxford.

HOFFMANN, T. A., and LADIK, J. (1961). *Cancer Res.* **21**, 474.

HOFFMANN, T. A., and LADIK, J. (1964). *J. Theoret. Biol.* **6**, 26.

HOFFMANN, T. A., and LADIK, J. (1965) *Advan. Chem. Phys.* **7**, 84.

HOROWITZ, N. H. (1945). *Proc. Natl. Acad. Sci. U.S.* **31**, 153.

HÜCKEL, E. (1931a). *Z. Physik* **70**, 204.

HÜCKEL, E. (1931b). *Z. Physik* **72**, 310.

HÜCKEL, E. (1932). *Z. Physik* **76**, 628.

HUGGINS, C. (1962). *In* "Horizons in Biochemistry" (M. Kasha and B. Pullman, eds.), p. 497. Academic Press, New York.

HUGGINS, M. L. (1962). *Am. Scientist* **50**, 485.

IVERSEN, S., and ARLEY, N. (1953). *Nature* **171**, 257.

JACQUIER, A., and DAUDEL, P. (1964). *Compt. Rend.* **258**, 5775.

JACOB, F., and MONOD, J. (1961). *J. Mol. Biol.* **3**, 318.

JEFFREYS, H. (1923). *Proc. London Math. Soc.* [2] **23**, 428.

JEHLE, H., PARKE, W. C., and SALYERS, A. A. (1964). *In* "Electronic Aspects of Biochemistry" (B. Pullman, ed.) p. 313. Academic Press, New York.

JOOP, N., and ZIMMERMANN, H. (1962). *Z. Elektrochem.* **66**, 440, 541.

KASHA, M. (1962a). *Proc. 11th Meeting Soc. Chim. 1961*, p. 38. Pergamon Press, Oxford.

KASHA, M. (1962b). *In* "Horizons in Biochemistry" (M. Kasha and B. Pullman, eds.), p. 583. Academic Press, New York.

KASHA, M., and PULLMAN, B. eds. (1962). "Horizons in Biochemistry," Albert Szent-Györgyi Dedicatory Volume. Academic Press, New York.

KASHA, M., EL-BAYOUMI, M. A., and RHODES, W. (1962). *Proc. 11th Meeting Soc. Chim. 1961*. Pergamon Press, Oxford.

KATRITZKY, A. R., and WARING, A. J. (1962). *J. Chem. Soc.*, p. 1540.

KATRITZKY, A. R., and WARING, A. J. (1963). *J. Chem. Soc.*, p. 3046.

KAUFMAN, H. E. (1963). *Proc. 3rd Intern. Congr. Chemotherapy, Stuttgart 1963.*

KAUFMAN, H. E., NESBURN, A. B., and MALONEY, E. D. (1962a). *Virology* **18**, 567.

KAUFMAN, H. E., NESBURN, A. B., and MALONEY, E. D. (1962b). *Arch. Ophthalmol.* **67**, 583.

KAUFMAN, H. E., MARTOLA, E. L., and DOHLMAN, C. (1962c). *Arch. Ophthalmol.* **68**, 235.

KENDREW, J. C. (1959). *Rev. Mod. Phys.* **31**, 94.

KENNER, G. W., REESE, C. B., and TODD, A. R. (1955). *J. Chem. Soc.*, p. 855.

KOCH, A. (1963). Unpublished.

KORNBERG, A. (1959a). *In* "A Symposium on Molecular Biology" (R. E. Zirkle, ed.), p. 31. Univ. of Chicago Press, Chicago, Illinois.

KORNBERG, A. (1959b). *Rev. Mod. Phys.* **31**, 200.

KORNBERG, A. (1962). "Enzymatic Synthesis of DNA." Wiley, New York.

KOSHLAND, D. E. Jr. (1962). *In* "Horizons in Biochemistry" (M. Kasha and B. Pullman, eds.), p. 265. Academic Press, New York.

KOZLOFF, L. M. (1959). *In* "A Symposium on Molecular Biology" (R. E. Zirkle, ed.), p. 178. Univ. of Chicago Press, Chicago, Illinois.

KRAMERS, H. A. (1926). *Z. Physik* **39**, 828.

KRAMERS, H. A. (1957). "Quantum Mechanics," pp. 60 and 61. North-Holland, Amsterdam.

LADIK, J. (1960). *Acta Phys. Acad. Sci. Hung.* **11**, 229.

LADIK, J. (1961). *Acta Phys. Acad. Sci. Hung.* **13**, 473.

LADIK, J. (1963a). *Acta Phys. Acad. Sci. Hung.* **15**, 287.

LADIK, J. (1963b). Preprint QB 8, Uppsala Quantum Chemistry Group.

LADIK, J. (1964a). *In* "Electronic Aspects of Biochemistry" (B. Pullman, ed.), p. 203. Academic Press, New York.

LADIK, J. (1964b). *J. Theoret. Biol.* **6**, 201.

LADIK, J., and APPEL, K. (1962). *Uppsala Quantum Chem. Group Tech. Notes* Nos. 78 and 79.

LADIK, J., and APPEL, K. (1964). *J. Chem. Phys.* **40**, 2470.

LADIK, J., and APPEL, K. (1965). Preprint QB 20, Uppsala Quantum Chemistry Group.

LADIK, J., and HOFFMANN, T. A. (1964a). *J. Theoret. Biol.* **6**, 26.

LADIK, J., and HOFFMANN, T. A. (1964b). *Biopolymers Symp.* **1**, 117.

LANDAU, L. D., and LIFSHITZ, E. M. (1958). "Quantum Mechanics." Pergamon Press, Oxford (see particularly pp. 171 and 305).

LÁNG, L., and LADIK, J. (1964). *Periodica Polytechnica Budapest*, p. 176.

LANGRIDGE, R., and GOMATOS, P. J. (1963). *Science* **141**, 694.

LENNARD-JONES, J., and POPLE, J. A. (1951). *Proc. Roy. Soc.* **A205**, 155.

LERMAN, L. S. (1963). *Proc. Natl. Acad. Sci. U.S.* **49**, 94.

LEVAN, A. (1959). *Svensk Naturvetenskap.* **12**, 212.

LEVINTHAL, C. (1959). *Rev. Mod. Phys.* **31**, 249.

LEVINTHAL, C., and CRANE, H. R. (1956). *Proc. Natl. Acad. Sci. U.S.* **42**, 436.

LIPMANN, F., HÜLSMANN, W. C., HARTMANN, G., BOMAN, H. G., and ACS, G. (1959). *J. Cellular Comp. Physiol.* **54**, Suppl. 1, 75.

LIPPINCOTT, E. R., FINCH, J. N., and SCHROEDER, R. (1959). *Proc. Intern. Symp. Hydrogen Bonding, Ljubljana 1957*, p.361. Pergamon Press, Oxford.

LIQUORI, A. M. (1963). Results reported at the Ravello Conference (unpublished).

LONGUET-HIGGINS, H. C., and COULSON, C. A. (1947). *Trans. Faraday Soc.* **43**, 87.

LONGUET-HIGGINS, H. C., and ZIMM, B. H. (1960). *J. Mol. Biol.* **2**, 1.

LÖWDIN, P.-O. (1950). *J. Chem. Phys.* **18**, 365.

LÖWDIN, P.-O. (1953). *J. Chem. Phys.* **21**, 496.

LÖWDIN, P.-O. (1954). *Proc. Symp. Mol. Phys. Nikko, Japan 1953*, p. 13. Maruzen, Japan.

LÖWDIN, P.-O. (1955). *Phys. Rev.* **97**, 1509.

LÖWDIN, P.-O. (1956). *Advan. Phys.* **5**, 1 (particularly pp. 49–56).

LÖWDIN, P.-O. (1959). *Advan. Chem. Phys.* **2**, 207.

LÖWDIN, P.-O. (1960). *Ann. Rev. Phys. Chem.* **11**, 107.

LÖWDIN, P.-O. (1962a). *J. Math. Phys.* **3**, 1171.

LÖWDIN, P.-O. (1962b). *J. Appl. Phys.* **33**, 251.

LÖWDIN, P.-O. (1962c). *Uppsala Quantum Chem. Group Tech. Notes* Nos. 64 and 65; *J. Mol. Spectr.* **14**, 112 and 119 (1964).

LÖWDIN, P.-O. (1962d). *J. Math. Phys.* **3**, 969.

LÖWDIN, P.-O. (1962e). *Uppsala Quantum Chem. Group Tech. Note* No. 83.

LÖWDIN, P.-O. (1962f). *Rev. Mod. Phys.* **34**, 520.

LÖWDIN, P.-O. (1963). *Rev. Mod. Phys.* **35**, 724.

LÖWDIN, P.-O. (1964). *Biopolymers Symp.* **1**, 161.

LÖWDIN ,P.-O. (1964b). *Biopolymers Symp.* **1**, 293.

LÖWDIN, P.-O. (1964c). *In* "Electronic Aspects of Biochemistry," p. 167 (B. Pullman, ed.). Academic Press, New York.

LÖWDIN, P.-O. (1965). *Mutation Research* **2**, 218.

MASON, R. (1960). *Cahiers Phys.* **124**, 473.

MAISIN, J. H. (1963). *Acta Unio Intern. Contra Cancrum* **19**, 453.

MENDEL, G. (1865). *Verhandl. Naturforsch. Vereins Brünn* **10**.

MESELSON, M., and STAHL, F. W. (1958). *Proc. Natl. Acad. Sci. U.S.* **44**, 671.

MILLER, R. L., and LYKOS, P. G. (1962). *Tetrahedron Letters* **11**, 493.

MILNE, W. E. (1930). *Phys. Rev.* **35**, 863.

MOTT, N. F., and SNEDDON, I. N. (1948). "Wave Mechanics and Its Applications," p. 15. Oxford Univ. Press (Clarendon), London and New York.

NAGATA, T. (1963). *Proc. Natl. Acad. Sci. U.S.* **49**, 551.

NAGATA, C., IMAMURA, A., SAITO, H., FUKUI, K.-I. (1963). *Gann* **54**, 109–117.

NIRENBERG, M. W., and MATTHAEI, J. H. (1961). *Proc. Natl. Acad. Sci. U.S.* **47**, 1588.

NIRENBERG, M. W., MATTHAEI, J. H., and JONES, O. W. (1962). *Proc. Natl. Acad. Sci. U.S.* **48**, 104.

NOMURA, M., HALL, B. D., and SPIEGELMAN, S. (1960). *J. Mol. Biol.* **2**, 306.

OCHOA, S. (1962). *In* "Horizons in Biochemistry" (M. Kasha and B. Pullman, eds.), p. 153. Academic Press, New York.

O'KONSKI, C. T. (1963). *Rev. Mod. Phys.* **35**, 732.

OPARIN, A. I. (1953). "The Origin of Life." Dover, New York.

OPARIN, A. I., ed. (1959). "International Symposium on the Origin of Life on the Earth." Pergamon Press, Oxford.

ORGEL, L. E. (1959). *Rev. Mod. Phys.* **31**, 100 (see the survey).

ORGEL, L. E. (1963). *Proc. Natl. Acad. Sci. U.S.* **49**, 517.

ORGEL, L. E., COTTRELL, T. L., DICK, W., and SUTTON, L. E. (1951). *Trans. Faraday Soc.* **47**, 113.

PAIGEN, K. (1962). *J. Theoret. Biol.* **3**, 268.

PAULING, L. (1939). "The Nature of the Chemical Bond." Cornell Univ. Press, Ithaca, New York (see p. 464 in 3rd ed., 1960).

PAULING, L., and COREY, R. B. (1951). *Proc. Natl. Acad. Sci. U.S.* **37**, 251.

PAULING, L., COREY, R. B., and BRANSON, H. R. (1951). *Proc. Natl. Acad. Sci. U.S.* **37**, 205.

PAUNCZ, R., DE HEER, J., and LÖWDIN, P.-O. (1962). *J. Chem. Phys.* **36**, 2247 and 2257.

PIMENTEL, G. C., and MCCLELLAN, A. L. (1960). "The Hydrogen Bond." Freeman, San Francisco, California.

PLATT, J. R. (1962). *In* "Horizons in Biochemistry" (M. Kasha and B. Pullman, eds.), p. 167. Academic Press, New York.

POLLARD, E. C. (1953). "The Physics of Viruses." Academic Press, New York.

POLLARD, E., and LEMKE, M. (1965). *Mutation Research* **2**, 214.

POPLE, J. A. (1951). *Proc. Roy. Soc.* **A205**, 163.

PRATT, D., and STENT, G. S. (1959). *Proc. Natl. Acad. Sci. U.S.* **45**, 1507.

PULLMAN, A., and PULLMAN, B. (1955a). *Advan. Cancer Res.* **3**, 117.

PULLMAN, A., and PULLMAN, B. (1955b). "La cancérisation par les substances chimiques et la structure moléculaire." Masson, Paris.

PULLMAN, A., and PULLMAN, B. (1959). *Biochim. Biophys. Acta* **36**, 343.

PULLMAN, A., and PULLMAN, B. (1960). *Rev. Mod. Phys.* **32**, 428.

PULLMAN, A., and PULLMAN, B. (1961a). *Nature* **189**, 725.

PULLMAN, A., and PULLMAN, B. (1961b). *J. Chim. Phys. (Paris)* p. 904.

PULLMAN, A., and PULLMAN, B. (1962a). *Nature* **196**, 228.

PULLMAN, A., and PULLMAN, B. (1962b). *In* "Horizons in Biochemistry" (M. Kasha and B. Pullman, eds.), p. 553. Academic Press, New York.

PULLMAN, A., and PULLMAN, B. (1962c). *Proc. 11th Meeting Soc. Chim. 1961*, p. 28. Pergamon Press, Oxford.

PULLMAN, B. (1962). *Compt. Rend.* **255**, 3255.

PULLMAN, B., and PULLMAN, A. (1962a). *Biochim. Biophys. Acta* **64**, 403.

PULLMAN, B., and PULLMAN, A. (1962b). *Nature* **196**, 1137.

PULLMAN, B., and PULLMAN, A. (1963). "Quantum Biochemistry." Wiley, New York.

QUASTLER, H., and ZUBAY, G. (1963). *J. Theoret. Biol.* **3**, 496.

REIN, R., and HARRIS, F. E. (1964a). *J. Chem. Phys.* **41**, 3393.

REIN, R., and HARRIS, F. E. (1964b). *Science* **146**, 649.

REIN, R., and HARRIS, F. E. (1965a). *J. Chem. Phys.* **42**, 2177.

REIN, R., and HARRIS, F. E. (1965b). Preprint QB 23, Uppsala Quantum Chemistry Group.

REIN, R., and LADIK, J. (1964). *J. Chem. Phys.* **40**, 2466.

RICH, A. (1958a). *Nature* **181**, 521.

RICH, A. (1958b). *Biochim. Biophys. Acta* **29**, 502.

RICH, A. (1959a). *Rev. Mod. Phys.* **31**, 50.

RICH, A. (1959b). *Rev. Mod. Phys.* **31**, 191.

RICH, A. (1959c). *In* "A Symposium on Molecular Biology" (R. E. Zirkle, ed.), p. 47. Univ. of Chicago Press, Chicago, Illinois.

RICH, A. (1960). *Proc. Natl. Acad. Sci. U.S.* **46**, 1044.

RICH, A. (1962). *In* "Horizons in Biochemistry" (M. Kasha and B. Pullman, eds.), p. 103. Academic Press, New York.

ROBERTS, R. B. (1962). *Proc. Natl. Acad. Sci. U.S.* **48**, 897.

SAGER, R., and RYAN, F. J. (1962). "Cell Heredity." Wiley, New York.

SCHRÖDINGER, E. (1945). "What is Life? The Physical Aspect of the Living Cell." Cambridge Univ. Press, London and New York.

SCHULL, W. J., ed. (1962). "Mutations. Second Conference on Genetics." Univ. of Michigan Press, Ann Arbor, Michigan.

SETLOW, R. B., and POLLARD, E. C. (1962). "Molecular Biophysics." Addison-Wesley, Reading, Massachusetts (see particularly p. 513).

SIMHA, R., and ZIMMERMAN, J. M. (1962). *J. Theoret. Biol.* **2**, 87.

SINGER, M. F., JONES, O. W., and NIRENBERG, M. W. (1963). *Proc. Natl. Acad. Sci. U.S.* **49**, 392.

SINSHEIMER, R. L. (1959). *J. Mol. Biol.* **1**, 43.

SINSHEIMER, R. L. (1962). *Proc. 11th Reunion Soc. Chim. Phys. 1961*, p. 110. Pergamon Press, Oxford.

SINSHEIMER, R. L., STARMAN, B., NAGLAR, C., and GUTHRIE, S. (1962). *J. Mol. Biol.* **4**, 142.

SINSHEIMER, R. L., BURTON, L., and KLEINSCHMIDT, A. K. (1963). *Science* **142**, 961.

SOKOLOV, N. D. (1959). *Proc. Intern. Symp. Hydrogen Bonding, Ljubljana 1957*, p. 385. Pergamon Press, Oxford.

SPEYER, J. F., LENGYEL, P., BASILIO, C., and OCHOA, S. (1962). *Proc. Natl. Acad. Sci. U.S.* **48**, 63, 282, 441.

STENT, G. S. (1958). *Advan. Virus Res.* **5**, 95.

STEVENS, A. (1961). *J. Biol. Chem.* **236**, PC43.

STRAUSS, B. S. (1960). "An Outline of Chemical Genetics." Saunders, Philadelphia, Pennsylvania.

SZENT-GYÖRGYI, A. (1960). "Introduction to Submolecular Biology." Academic Press, New York.

SZENT-GYÖRGYI, A., HEGYELI, A., and MCLAUGHLIN, J. A. (1963). *Science* **140**, 1391.

SZILARD, L. (1959). *Proc. Natl. Acad. Sci. U.S.* **45**, 30.

SZILARD, L. (1960). *Nature* **186**, 649.

TAYLOR, J. H. (1959). *In* "A Symposium on Molecular Biology" (R. E. Zirkle, ed.), p. 304. Univ. of Chicago Press, Chicago, Illinois.

TAYLOR, J. H., WOODS, P. S., and HUGHES, W. L. (1957). *Proc. Natl. Acad. Sci. U.S.* **43**, 122.

TESSMAN, I. (1959). *Virology* **9**, 375.

TSCHERMAK, E. (1900). *Ber. Deut. Botan. Ges.*

TSUGITA, A., and FRAENKEL-CONRAT, H. (1960). *Proc. Natl. Acad. Sci. U.S.* **46**, 636.

TSUGITA, A., and FRAENKEL-CONRAT, H. (1962). *J. Mol. Biol.* **4**, 73.

VEILLARD, A., and PULLMAN, B. (1963). *J. Theoret. Biol.* **4**, 37.

VIELMETTER, W., and WIEDER, C. M. (1959). *Z. Naturforsch.* **14b**, 312.

VOLKIN, E. (1960). *Proc. Natl. Acad. Sci. U.S.* **46**, 1336.

VOLKIN, E., and ASTRACHAN, L. (1956). *Virology* **2**, 149.

VOLKIN, E., ASTRACHAN, L., and COUNTRYMAN, J. L. (1958). *Virology* **6**, 545.

WALPOLE, A. L. (1959). *Ciba Found. Symp., Carcinogenesis, Mechanisms Action*, p. 41.

WATSON, J. D., and CRICK, F. H. C. (1953). *Nature* **171**, 737, 964.

WENTZEL, G. (1926). *Z. Physik* **38**, 518.

WHELAND, G. W. (1942). *J. Am. Chem. Soc.* **64**, 902.

WHELAND, G. W., and PAULING, L. (1939). *J. Am. Chem. Soc.* **57**, 2091.

WILKINS, M. H. F. (1957). 'The Structure of the Nucleic Acids and Their Role in Protein Synthesis." Cambridge Univ. Press, London and New York.

WILKINS, M. H. F. (1961). *In* "Biological Structure and Function," Symposium (T. W. Goodwin and O. Lindberg, eds.), p. 13. Academic Press, New York.

WILKINS, M. H. F., STOKES, A. R., and WILSON, H. R. (1953a). *Nature* **171**, 738.

WILKINS, M. H. F., SEEDS, W. E., STOKES, A. R., and WILSON, H. R. (1953b). *Nature* **172**, 759.

WILLIAMS, R. C. (1959). *Rev. Mod. Phys.* **31**, 233.

WOESE, C. R. (1961a). *Biochem. Biophys. Res. Commun.* **5**, 88.

WOESE, C. R. (1961b). *Nature* **190**, 697.

WOESE, C. R. (1962). *Nature* **194**, 1114.

WITTMANN, H. G. (1961). *Naturwiss.* **48**, 730.

WOLSTENHOLME, G. E. W., and O'CONNOR, M., eds. (1959). *Ciba Found. Symp., Carcinogenesis, Mechanisms Action.*

WOOD, W. B., and BERG, P. (1962). *Proc. Natl. Acad. Sci. U.S.* **48**, 94.

YAJIMA, T., and ESAKI, L. (1958). *J. Phys. Soc. Japan* **13**, 1281.

YCAS, M. (1959). *In* "A Symposium on Molecular Biology" (R. E. Zirkle, ed.), p. 115. Univ. of Chicago Press, Chicago, Illinois.

YCAS, M. (1960). *Nature* **188**, 209.

YCAS, M., and VINCENT, W. S. (1960). *Proc. Natl. Acad. Sci. U.S.* **46**, 804.

YOSHIKAWA, H., and SUEOKA, N. (1963a). *Proc. Natl. Acad. Sci. U.S.* **49**, 559.

YOSHIKAWA, H., and SUEOKA, N. (1963b). *Proc. Natl. Acad. Sci. U.S.* **49**, 806.

ZIMMERMANN, H. (1959). *Z. Elektrochem.* **63**, 601.

ZIMMERMANN, H. (1961). *Z. Elektrochem.* **65**, 821.

ZIRKLE, R. E., ed. (1959). "A Symposium on Molecular Biology." Univ. of Chicago Press, Chicago, Illinois.

ZUBAY, G. (1958a). *Nature* **182**, 112.

ZUBAY, G. (1958b). *Nature* **182**, 1290.

ZUBAY, G. (1962). *Proc. Natl. Acad. Sci. U.S.* **48**, 456.

ZUBAY, G. (1963). *Science* **140**, 1092.

ZUBAY, G., and QUASTLER, H. (1962). *Proc. Natl. Acad. Sci. U.S.* **48**, 461.

ZUCKERKANDL, E., and PAULING, L. (1962). *In* "Horizons in Biochemistry" (M. Kasha and B. Pullman, eds.), p. 189. Academic Press, New York.

Author Index

Numbers in italics refer to pages on which the complete references are listed.

Subject Index