THE DESIGN OF BIOLOGICALLY ACTIVE POLYPEPTIDES

Author: **Barry Robson**
Department of Biochemistry
University of Manchester
Manchester, England

Referee: Jean Garnier
Laboratory of Biochemistry
University of Paris
Paris, France

I. INTRODUCTION

To "design" a molecule means to predict that covalent structure which, when synthesized, will have certain required properties.[1] Since the molecule does not exist, this is primarily a problem in theoretical chemistry.

Here we are explicitly interested in polypeptides (including modified peptides, oligopeptides, and proteins) and, thus, in predicting the sequence of amino acid residues which will have the required properties. However, even natural proteins would be largely confined to simple catalytic functions of the hydrolytic type were it not for their association with cofactors including inorganic ions, and the designer is certainly free to introduce new chemical groups. From the functional point of view, polypeptides can be divided into those which have their function (1) in vitro, such as artificial enzymes, and (2) in vivo, such as pharmaceutical agents. In the future, it may also be possible to discuss the role of artificial or modified natural proteins as microminiaturized machinery which interconverts, harvests, or measures (1) light, (2) chemical energy, (3) mechanical energy, and (4) electrical energy, or which handles information in microscopic computer "chips" by electron transport.[2] Although we would be asking proteins to do no more than exhibit the kind of function they possess in living cells, such discussion would be premature, and we will be principally concerned with artificial enzymes and pharmaceutical agents.

So far, most effort has been in relation to the pharmaceutical aspect. The impetus is largely financial since methods of developing an initial *leader compound* are already well established, but are largely based on the trial-and-error testing of a large number of derivatives which could take more than 60 man-years and cost more than $40 million.[3] Even if design procedures are not completely reliable, they would certainly provide a cost-effective screening procedure by eliminating unlikely compounds. Even here success has been limited to small peptide derivatives such as penicillins and nitroso-ureas, though the techniques used in design have frequently been applied to deduce the shape and mode of action of neuropeptides and oligopeptide hormones.[4-6,127]

"Design" of polypeptides does not simply imply the passive elimination of those derivatives which do not lead to the required function, however. It implies the shaping of a conceptual model towards the required goal. This is helped by the degeneracy of both amino acid sequence with respect to conformation and of conformation with respect to function. We know from divergent evolution that many proteins of similar conformation and function can have very different sequences, and from convergent evolution that proteins with similar functions can have different conformations (e.g., subtilisin and a serine protease). The designer is free to select from many sequences which may have the appropriate physicochemical properties, and these need not be natural-looking sequences.[1]

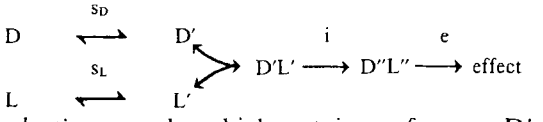
On the other hand, the designer must generally adhere to certain fundamental design stages.[1] By *function analysis* he decides what kind of functions he requires and those which he does not want, e.g., toxicity in the pharmaceutical context. He must quantify these notions so as to optimize the desired effects and minimize the undesired effects. He must identify any other molecular species over which he has no direct control (e.g., a drug receptor), and this may necessitate further experiments, e.g., binding studies using different drug analogs to identify the *pharmacophore* (form of the receptor site or the essential complementary features of the drug). By *activity design* he choses any required catalytic groups and their arrangement in space, and by *specificity design* he choses binding groups and their arrangement in space. By *scaffold design* he attempts to find a chemical structure which places the catalytic and binding groups in the required spatial arrangement. Because each residue may have five or more rotatable bonds, the latter is arguably the most difficult stage and will be given particular attention.

Because scaffold design is a complex problem, it is tempting to consider the use of molecular species, other than polypeptides, with a rigid shape determined by covalent bonds. For example, morphine has such a rigid shape and yet is an excellent functional analog of the polypeptide enkephalins. Nonetheless, the designer is tempted by the great variety of form and function of natural polypeptides, and in many cases has detailed structural information from which to draw useful "rules of thumb". Still more important, it will be possible to exploit genetic engineering so that bacteria can be adapted to produce the designed polypeptides in quantity.

II. FUNCTION ANALYSIS

This is the most difficult design stage to discuss in general and formal terms, since its role is to define the design problem more precisely in any given context. Nonetheless, it plays such a fundamental role in the design procedure that it must be considered.

In essence, the designer will always want to optimize some set of *goal properties* as a function of the chemical structure. These properties could be assigned different weightings, since all properties are not generally of equal importance, and undesirable properties can be assigned a negative weighting so that these will be minimized. This set of goal properties will depend on the system as a whole, including a drug receptor in pharmaceutical context or the substrate in an enzyme design context. However, even the simplest system presents a number of problems. Let the peptide D (e.g., drug or substrate) bind to just one other molecular ligand L (e.g., the receptor or enzyme). Then the general representation is



Here, s_D and s_L are *selection steps* by which certain conformers D′ and L′ are selected from the equilibrium population for binding, i is an *induction step* in which a certain conformation is induced on formation of complex D′L′, and e is the *function effector* step (typically rate limiting).

If L is a natural receptor, the ability to undergo the transition from D′L′ to D″L″ will be required to initiate the "effect", i.e., the response of the cell. Conversely, if L is an artificial enzyme and D the substrate, the same conformational transition may be involved in dynamic aspects of catalysis, or perhaps in control.

The selection steps are also important. However, the effective conformation of a drug will not generally be D′ but D″ which may be of higher free energy though it will not generally be higher than the free energy of formation of the D′L′ or D″L″ complex (whichever is of lower free energy). Conversely, the overclever designer may seek to

enhance the probability of D″ and discover that his molecule prohibits the reaction for formation of the D′L′ complex. Similarly, if he seeks to enhance D′ over D, he has neglected the possibility that selection of the less stable D′ from the equilibrium D-D′ population is an important source of strain to be used in the function. It follows that the only safe bet is to design for the stability of D,D′ and D″, and to tidy up their relative probabilities in solution in a later refinement stage.

A. Quantification of Goal Properties

In the very simplest case, we can assume only one conformer for dissociated species, i.e., D′ and L′. Then, we are largely concerned with the formation (or dissociation) constant for the process D′ + L′ = D″L″, and this must be related to the biological response. Beddell et al.[4] analyzed the relationships between enkephalin derivatives and the biological response, thus, laying the foundation for the design of "improved" enkephalins. They sought chemical modifications of leucine-enkephalin which would inhibit, to markedly different extents, the neurally invoked contraction of the isolated vas deferens of the mouse. The degree of contraction Y in response to the neurally invoked contraction was simply the difference between the initial and final lengths. To arrive at some measure Q of potency which is independent of enkephalin concentration and preparation they used the relation suggested by Young:[7]

$$Y = \frac{C - mQX}{1 + QX} \qquad (1)$$

where X = enkephalin dose, C = contraction at zero dose, and m = the asymptotic contraction at infinite dose. The potency factor Q is the reciprocal of the widely used $C_{50}M$, i.e., the dose X giving halfway between maximum and minimum response, but here in terms of inhibition of neural activity. In general, if $C_{100f}M$ is the molar concentration required to give fractional response f (above, $f = [C - Y]/[C - m]$), then at equilibrium

$$C_{100f}M = \frac{f}{1 - f} K \qquad (2)$$

and, hence,

$$C_{50}M = K \qquad (3)$$

where K is the dissociation constant for the process $D''L'' \rightleftharpoons D' + L'$ with biologically inactive D′L′ as intermediate. Exactly the same situation arises in the Michaelis-Menten treatment of enzyme reactions, where $f = v_o/(V_{max} - v_o)$ and K is replaced by Km, the ratio $(D') \cdot (L')/(D''L'') = (k_{-1} + k_2)/k_{+1}$ *under steady state.* The "effect" v_o is, of course, proportional to (D″L″) and equals k_2 (D″L″); $C_{50}M$ corresponds to Km.

In using simple models it seems reasonable to suppose that $C_{50}M$ is independent of the concentration of the species with which the designed molecule forms the complex, but this is not always so in analogous situations for enzymes (where Km is dependent on the concentration of enzyme), and even less often the case in the pharmacological context (where $C_{50}M$ is dependent on the concentration of receptor sites). This arises when the enzyme concentration or receptor site concentration is large or the dissociation constant K (or Km) is small, so that the concentration of free drug or receptor is not equivalent to the controlled concentration added to the study but is significantly depleted by complex formation. This is irksome because we cannot maximize Q (minimize $C_{50}M = Q^{-1}$) directly, but must determine K and minimize that as a function of the chemical

structure of the design molecule. This variation in $C_{50}M$ has been considered by Chaplin[8] and Robson,[9] amongst others, in relation to enzymes and either approximations[10] or simplified mathematical descriptions[9] sought. Following the reasoning of Robson,[9] it is easy to show that, in the design of pharmaceutical agents,

$$C_{50}M = K + \tfrac{1}{2}(L) \qquad (4)$$

where (L) is the concentration of receptor site and K the dissociation constant.

Robson[9] has also demonstrated and analyzed another very common complication, when the substrate of an enzyme is conformationally flexible and represents an equilibrium population of forms only some of which are selected for binding. In the case of two conformers D and D′ as in scheme I, with k_D the rate constant for formation of D from D′ and $k_{D'}$ for the formation of D′ from D, we obtain

$$C_{50}M = K_m\left(1 + \frac{k_D}{k_{D'}}\right) + \tfrac{1}{2}(L)\left(1 + \frac{k_{cat}}{k_{D'}}\right) \qquad (5)$$

Here $C_{50}M$ is not only dependent on the enzyme concentration of (L) but also on the rate constants for conversion between D and D′ (and, hence, on the energy barrier for D and D′ interconversion). Note also that the rate constant $k_{D'}$ will cause the receptor concentration term to dominate the value of $C_{50}M$ if k_D is much smaller than k_{cat} (the limiting catalytic rate constant for expression of the effect of the complex; Scheme I), and that high concentrations of L or low values of K_M are no longer essential in order that $C_{50}M$ be strongly dependent on (L). Indeed, when $k_D < k_{cat} > k_{D'}$ and studies are genuinely carried out under steady-state conditions, K_m may make little contribution to $C_{50}M$. In the case of a drug-receptor interaction without catalytic transformation, we simply omit the term $(1 + k_{cat}/k_D)$, i.e., $k_{cat} = 0$. K_m will respond to the dissociation constant for the process $D''L'' \rightarrow D' + L'$, while $(1 + k_D/k_{D'})$ introduces the contribution of the transition $D \rightleftharpoons D'$. Behavior in the more classical, except that with low stability for D′ (i.e., $k_{D'} < k_D$), the effect of receptor concentration will actually be decreased as $K(1 + k_D/k_{D'})$ dominates.

Clearly, the system must be well understood before it can be assumed that $C_{50}M$ (and thus $Q = [C_{50}M]^{-1}$) relates simply to a single effective parameter K as implied by Equation 3. One should ideally carry out experiments and, using graphic methods, determine K and any other relevant parameters such as k_D and $k_{D'}$. These experiments will depend on determining $C_{50}M$ at different concentrations of receptor in the drug case, or enzyme in the classic enzymological case of determining K_m.

It is important to note that Equations 2 to 5 applied to drug systems assume a direct relationship between the concentration of complex and the biological effect, as for a simple enzyme system where the rate of product formation is proportional to the concentration of complex. If this assumption were unwarranted, we have a further complication. After all, there may be many complex events between the activation of the receptor and the final biological response when considering drug action. Beddell et al.[4] examined many opiate and enkephalin analogs and found a close relation between concentrations giving half maximum biological effect and half maximum binding of the drugs to the receptors. Nonetheless, in log-log plots of one measure against the other, there was a departure of the slope by some 20% from the expected value of unity.

In the simplest receptor models treating this problem.[11-13] L can be regarded as a complex formed by an association between an enzyme E, to be activated, and the receptor R, such that, for example, the complex D″L″ may be written as D″R″E. The steps leading to "effect" in Scheme I can then be replaced by

$$D'R'E \rightarrow D''R''E \rightarrow D''R''E^* \qquad II$$

where E* is the active form of the enzyme.[11] This is typified by the aspartate transcarbamylase system where D′ represents the activator ATP or CTP. However, there are also cases (e.g., cAMP-dependent protein kinase) where E must first dissociate from D″R″ prior to adopting its active form E*, and probable cases (e.g., for the insulin receptor[14]) in which E is normally dissociated from D′R′ and must first associate with it to produce D′R′E. For α-receptor adenylate kinase,[12] both these may apply: E must first associate with D′R′ and then E* is subsequently released. The last case also typifies the complexity of subsequent steps which E* may have to initiate before the final measured effect is obtained. The cAMP released initiates a chain of three enzyme steps before glycogen is converted to glucose-1-P. Thus, the assumption that Q^{-1} is simply related to the dissociation constant K for the DL complex is a dangerous one, and for the more quantitative studies discussed below experimental estimates of K from binding studies are desirable.*

B. Relation of the Goal Property to Physicochemical Properties

If one is to optimize some goal property as a function of chemical structure, then one must ask how the goal property will vary with the physicochemical properties which each chemical structure implies. One requires a predictor equation relating Q, for example, to the properties; such an equation can be derived experimentally (on the basis of observed relationships), theoretically, or by a combination of both. Typically, with properties p,q,r, . . . one attempts to fit a linear equation[15] of the form

$$Y = a + bp + cq + dr + \ldots \qquad (6)$$

where Y is some goal property such as Q and where coefficients a,b,c,d, . . . emerge from the equation of best fit. Other approaches include discriminant analysis[16] and learning machines.[17]

Earlier reviewers[6] have noted the tendancy for workers to consider a large variety of potentially useful properties, most of which have not been independent. More recently, attention has focused on electronic, steric,[18] and solvent-dependent properties.[19] However, many workers have neglected all but one of these properties, considering that one as being of particular relevance. For example, Montgomery et al. correlated the tumor-delaying potencies of nitrosoureas ($R.NHCO[NO]CH_2CH_2Cl$) with the octanol-water partition coefficients (P) of the whole molecule[19] for different groups R. For the above linear equation they obtained a good fit[19] with

$$-\log(C) = 1.23 + 0.14 \log (P) - 0.08 \log^2(P)$$

where C is a quantity analogous to $C_{50}M$ and, hence, $-\log(C)$ is analogous to $\log(Q)$. This use of logarithmic terms is very frequent and often reflects the fact that RTlnk represents a free energy of drug-receptor complex conformation to which potential energies due to steric and electronic factors, and free energies due to solvent effects, may

* Provided that there is a porportional relationship of biological response to the concentration of D″L″, however, the use of $C_{50}M$ has one important advantage. From Scheme I we find that the biological response is proportional to (D″L″), while the affinity constant is proportional to (D′L′) + (D″L″), the total concentration of complexed species. In combined experimental and theoretical studies, or in calculation of the experimental quantities from theoretical principles, calculation of the affinity or dissociation constant requires consideration of the extra species D′L′.

contribute more or less additively. In this context, also, the classic Free-Wilson[20] assumption of a linear relationship between net and component group properties seems well founded.

For true peptide systems, however, it has been the steric factor which has been treated as of greatest importance. Presumably, this is because the steric aspect is emphasized in the "lock and key" fit concept of drug-receptor action. Marshall and Bosshard[21] advocated the synthesis of analogs of biologically active polypeptides, substituting various residues by those such as proline and dimethyl glycine which have reduced conformational freedom. If activity is low for such an analog, the active conformer should belong to the newly excluded part of conformational space. If it stays the same or rises, it must belong to the part of conformational space accessible to the analog. The possibility of a direct inhibition of binding by a large group, without necessarily changing the conformation, must, of course, be born in mind. Monahan et al.[5] used a similar rational in replacing an internal glycine residue of leutinizing hormone-releasing factor (LHRF) by L-alanine, then D-alanine. The D-alanine promoted activity, while L-alanine inhibited it, suggesting that the glycine was adopting a conformation characteristic of a type-II reverse turn conformation.[22] Bedell et al.[4] have used a similar approach, substituting glycine, L-alanine, and D-alanine at a variety of different locations in leucine enkephalin. They also deduced the possibility of a type of reverse turn structure, as also suggested in a comparison of enkephalin and its rigid morphine analog[23] and in a conformational energy calculation.[24]

Although such studies seem useful, from a more general theoretical viewpoint they appear to fall far short of a complete treatment.[25] After all, even neglecting induction, the proper treatment would require estimation of the free energies of the D'L' complex and of the isolated (or infinitely separated) species D and L:

$$RT\ln k = \Delta F_{D'L'} - \Delta F_D - \Delta F_L \tag{7}$$

or, alternatively,

$$RT\ln k = \Delta F_{assoc} + \Delta F_{select,D} + \Delta F_{select,L} \tag{8}$$

where the free energy of association between D' and L' is

$$\Delta F_{assoc} = \Delta F_{D'L'} - \Delta F_{D'} - \Delta F_{L'} \tag{9}$$

and the free energies of the selection steps are

$$\Delta F_{D,select} = \Delta F_{D'} - \Delta F_D \tag{10}$$

$$\Delta F_{L,select} = \Delta F_{L'} - \Delta F_L \tag{11}$$

The free energy ΔF_X of each species with its solvent is given by[25-27]

$$\Delta F_X = -RT \int_q \int_p \left\{ \exp - [KE(p,q) + U(q)]/RT \right\} \partial p \cdot \partial q \tag{12}$$

where KE and U are the kinetic and potential energy contributions of the nuclei as a function of their positional coordinates q and momenta p (an additive constant dependent on the distinguishability[27] of the separate particles of the system can be neglected in considering relative free energies of D'L' and D + L at infinite separation). In contrast, the approaches (References 5, 21, 22, but not 23) of the previous paragraph

appear to consider only the enthalpy change $\Delta H_{D,select}$ of the selection process D to D′, and treat this enthalpy qualitatively in terms of allowed or disallowed conformers.

C. The Pharmacophore Problem

The reason for neglect of L′,L′ and L is, of course, that the structure of receptor L is unknown; this is the classic problem of the *pharmacophore;* the structure of some qualitative representation of the receptor site for drug binding or of the drug which is complementary to that site must be deduced indirectly. The method of Marshall and Bosshard[21] and related methods do, nonetheless, permit this deduction in a way which may be reexpressed as follows. If each calculated free energy contribution ΔF_i for conformer i of dissociated drug is transformed to the corresponding statistical weight $\omega_i = \exp(-F_i/RT)$, then knowing K^{-1} experimentally one can set up a series of equations, each equation for one analog A:

$$K_A^{-1} = \sum_i a_i(A)\,\omega_i(A) / \sum_i \omega_i(A) \qquad (13)$$

The coefficients $a_i(A)$ for each analog A are the unknowns and represent, appropriately, normalized statistical weights for the association and induction process between D′ and L′. K_A is again a dissociation constant of analog A. If L is not equivalent to L′, the statistical weight for the L to L′ transition is also absorbed into the coefficients. One then minimizes $\left\{K_A^{-1} - \sum_i a_i(A)\,\omega_i(A) \sum_i \omega_i(A)\right\}^2$ as a function of the a_i, considering the latter as invariant of A. This may be facilitated by the assumption, implicit in the Marshall-Bosshard procedure,[21] that all a_i but one (or one set of closely related conformers) for any analog are zero, which is to assume that only one conformer has significant binding. Then, nonzero a_i values can be found by inspection.* Whether or not this simplification is made, the conformer associated with the highest value a_i is the "active" binding conformer. Clearly, the method works best for a large number of analogs with a broad range of dissociation constants, and care must be taken with group substitutions which can inhibit association rather than alter the conformational preferences of the drug (when a_i depends on A).

Specific treatment of association and induction steps appear in the increasingly popular "zipper"[28] or "dynamic receptor"[29] models of both drug and enzyme action. Burt et al.[30] proposed that enkephalin first binds to the receptor via its tyrosamine moiety, and subsequently an induction step drives the C-terminal end of enkephalin into the required conformation for the active D″L″ complex. This deduction was based on the similarity between the most stable enkephalin conformation found by calculation, and the structure of the relatively rigid morphine analog, except for the disposition of the C-terminus. However, on the available data this could have equally well been envisaged as a selection step, since the active conformation is also likely to be accessible for dissociated enkephalin, albeit a higher energy conformer. There are kinetic differences, however. In enzyme systems the selection step may enhance the dependence of the apparent K_m on the

* In the most convenient case, which is to say for a binding mechanism leading to a simple analysis by this method, a plot of log K vs. log (statistical weight of active conformer/sum of statistical weights of all conformers), each point representing a different analog A, will be linear. It will be linear only for the active conformer, which is, thus, identified. Such a binding mechanism is when the receptor is floppy and easily adopts any conformation consistent with each drug conformer, but with only one of these consistent conformations representing the active one D″L″. Chemical changes altering the ease of association will cause departures from the linear plot. Ideally, use of $\log(C_{50}M)$ is better than log K, since the statistical weight of the initial complex D′L′ plus the statistical weight of the active complex D″L″ contributes to the theoretical description of the experimental dissociation constant.

receptor concentration,[9] while the induction step can be absorbed into classical, Michaelis-Menten-type kinetics. The induction step as described in the model of Burt et al.[30] would also imply higher rate constants for complex formation, since the required conformation would be steered into the required form rather than achieving it by chance,[11] a possible saving in activation entropy of 8 kcal mol^{-1}.

For many purposes, nonetheless, it may be sufficient to neglect specific treatment and to concentrate on improved methods for determining statistical weights ω. Finn and Robson and Robson et al.,[32] thus, determined the stable and metastable conformers of thyroid hormone releasing factor (TRF) and analogs by energy minimization as a function of dihedral bond angles, from a variety of starting conformations. One important effect of the solvent was included via the reaction field of Onsanger.[33] However, since the binding process actually takes account of drug surface features, not the drug dihedral angles, these conformers were classified on the basis of a "surface" feature which adequately characterized the conformers, namely, the relative disposition of the sidechain rings of pyroglutamic acid, histidine, and proline. Since the TRF sequence is pyroGlu.His.Pro.NH_2, this clearly provides a fairly comprehensive summary of the molecule. TRF and analogs all had some or all of the following conformers (see Figure 2): a propellar P form where the three rings are equatorially displayed like blades of a three-bladed propellar, a cup C form where the three rings approach each other like the cupped petals of a flower, and three ring stacked Y forms in which each possible pair of the three rings stacked together with the remaining ring jutting out in the opposite direction. A high statistical weight for the P conformer was found consistent with the binding and activity data.

Although the conformer classes were constructed on the basis of the distances between ring centers and involved a cluster analysis of points in the space of the three interring distances to rationalize the choice, it could still be argued that this introduces a subjective element. After all, certain other distances could be equally important. Crippen[34] has defined a more general procedure in which all distances between all chemical groups are given equal prior significance, upper and lower bounds subsequently being deduced on the basis of comparison with the binding data and stereochemical reasoning. This technique was applied to the study of inhibitors of serine proteases. The approach differs somewhat in that the use of the distance matrix dominates the calculation; accurate free energy estimates were not exploited though they could, of course, be introduced in a more refined stage of the study. It may well be argued that more refined calculations of the conformational free energy are not justified, since the receptor structure and the precise interactions with the polypeptide drug are unknown and introduce still greater uncertainty. The binding free energies for many polypeptide hormones with receptors tend to be lower than -10 kcal mol^{-1}, a considerably stronger contribution than a typical enzyme-substrate association at about -5 kcal mol^{-1} and undoubtedly a very significant contribution.

In any event, the objection would certainly vanish if the receptor structure were known to the extent that its interactions with the drug could be calculated. By using the above techniques to deduce the active drug conformer (that with the maximum a_i) one can, of course, make some inference about the complimentary receptor site, but rarely to provide a description sufficiently precise to allow detailed calculations of the binding energy for any new polypeptide drug analog, which is the essence of real design. Perhaps the use of genetic analogs of insulin[35] and glucagon[36] hormones is bringing us close to this ideal, but much remains to be done. An alternative approach would be a good prediction of the receptor site on predominantly theoretical grounds. With a view of designing anticlotting agents, a structure of the thrombin B chain has been proposed[1] on the basis of its amino acid sequence and knowledge of the conformation of the homologous elastase. This involved fitting the thrombin structure to elastase by minimizing the root-mean squared

distance of corresponding atoms (allowing for insertion regions), and then minimizing the energy of the thrombin as a function of bond dihedral angles. Another homolog, trypsin, was used as a control study; the structure is known and in quite tolerable agreement with the predicted trypsin structure (2.4 Å rms between calculated and observed C^{α} coordinates). Again, however, much remains to be done, and this approach is limited to cases where the amino acid sequence and a homologous protein of known structure are available.

III. ACTIVITY DESIGN

The term "activity" here is borrowed from the field of enzymology; it refers to catalytic functions, not biological activity in the general sense. Thus, this design stage is primarily of interest to artificial enzyme design, and involves choosing catalytic groups and their arrangement in space. The question of how to achieve that arrangement in space arises at a later stage of design. Nonetheless, some of the theoretical aspects are important to pharmacologists when a drug is chemically modified by the receptor. Boyd[37] has defined two factors required for *recognition* of a β-lactam antibiotic by its receptor, namely, the lactam carboxyl group and the ring carboxyl group. The *potency,* however, depends on the energy difference between the ground and transition states, so that lnQ is a linear function of terms dependent on this energy difference, estimated by quantum mechanical calculation.

Quantum mechanical calculations like those used by Boyd[37] are of increasing importance in enzyme design, but, previously, standard text books of chemical catalysis were exploited as directories of known options. For example, Dhar and colleagues[38] argued that glycosidic activity should result if glucosidic oxygen atoms could be protonated by an appropriately placed unionized carboxyl group, and if the resulting carbonium ion could be stabilized by an ionized carboxyl group. Gutte et al.[39] designed a nuclease with emphasis on the known catalytic role of histidine.

One of the first quantum mechanical studies of enzyme action was that of Warshel and Levitt.[40] However, these studies went beyond the idea of simply calculating excited states and the idea of the electrostatic field contributed by the enzyme became increasingly more important. Warshel and Weiss[41] have argued that a dominant catalytic effect is the stabilization of intermediate ionic configurations and alteration of intrinsic pK values of groups. Allen[42] has examined a number of enzymes and calculated the fields in the vicinity of the active site. It is now seen as important to plot the field generated by the molecule, whereas earlier workers like Dhar and Gutte saw the electrostatic effect as predominantly short range. This is to say they considered the electrostatic effect from roughly the same point of view as one uses the hard sphere approximation for van der Waals' interactions. Like the latter approximation, the short-range view is easy to use and often effective, but suffers from the inability to calculate degrees of effects and gives a very qualitative picture.

IV. SPECIFICITY DESIGN

The binding of high energy intermediates to an enzyme (a question of enzymic activity) cannot always be wholly divorced from binding of the ground state (which relates to specificity). Nonetheless, natural enzymes usually have groups which seem more concerned with promoting binding of one substrate species, and decreasing that of other substrate species, than catalysis per se. In the calculation of the thrombin active site[1] in order to design inhibitors, a problem was that the short associated A chain could greatly influence specificity, yet no known extensive homolog of this A chain was available to aid the calculation of its structure. The distinction between catalytic and binding *sites* is

particularly obvious in molecules with large substrates, such as the serine protease, where binding and catalytic sites may be quite far apart on the enzyme surface.

In the current status of the art of enzyme design, enhancing the binding of the required substrate is seen as much more important than decreasing the binding of other substrate species. Dhar and colleagues[43] reasoned that in designing a polypeptide, to bind acetylcholine, one may employ a carboxyl group to attract the positively charged quaternary ammonium moiety, and an amino or hydroxyl group to hydrogen bond to the ester carboxyl group. In seeking to enhance binding of nucleotides to their artificial protein, Gutte et al.[39] considered hydrogen bonding and hydrophobic interactions. In the development of their glycosidase, Dhar et al.[38] first tried a binding site of alanine and phenylalanine sidechains, but on the grounds that this site did not seem sufficiently hydrophobic, replaced one of the alanine residues by a further phenylalanine; this significantly improved the result.

Obviously, the consideration of the electrostatic field is as important to the stabilization of the enzyme-substrate complex as it is to the stabilization of the excited state of the substrate in the complex. Allen[42] has demonstrated that the dissociation constant K, the approach of the substrate to the enzyme, its preferred initial point of binding, and, hence, the induction step, are all dependent on the spatial distributions of the electrostatic field. Again, however, the näive view often seems sufficient in the current status of the art. Electrostatic effects on binding have traditionally been treated in terms of placing negative or partial negative binding site charges adjacent to those on the binding molecule[38,43] and this has lead to quite reasonable conclusions.

V. SCAFFOLD DESIGN

The "scaffold" is the molecular structure to which the catalytic and binding groups are attached in order to give them the required spatial arrangement for catalysis and specificity. Whereas in vivo there are fewer choices because of the typically small size of molecule involved and the possible need to pay some attention to questions of absorption, transport, degradation, modification, and toxicity, almost any structure of support can be used in vitro irrespective of the unbiological appearance. Thus, Chakravarty et al.[38] used an α-helical support of ten residues to support groups with glycosidase activity, and of five residues to produce acetylcholine binding. Such a scaffold may not be a universal choice in that short helices are notoriously unstable as a consequence of the high free energy which must be invested in order to nucleate the helix. Only for long helices can one guarantee that favorable energy of atomic packing and hydrogen bond formation pays off this initial capital investment, except in globular protein where favorable interactions between structural components "nationalize" the industry of helix formation.

With this in mind Gutte et al.[39] attempted a nuclease with both an α-helical part and a hairpin of pleated sheet, though not involving the same arrangements as encountered in the nucleotide domains of natural proteins.

It is clear that these designers cannot *enforce* the helix or pleated sheet components directly, rather they construct amino acid sequences *most likely* to give rise to such structures. This question is, of course, central to the design problem and is considered below. However, it is perfectly feasible to design very unbiological scaffolds with a very high chance of giving the required arrangement. A small synthetic analog of the active site of hemoglobin has been constructed[44] using the propionic sidechains of protohaem IX as supports for the two histidine residues which chelate the Fe ion above and below the ring. One of these histidines was also linked via glycine to a large polyethylene glycol support. This design was extremely successful in that reversible oxygen binding and characteristic hemoglobin spectra were obtained, though this success obviously owes

much to the specific coordination stereochemistry of Fe. The other artificial enzyme systems described above had no such help and the real degree of success demands more detailed scrutiny.

VI. ASSESSMENT OF SUCCESS

At face value, the results for artificial enzymes seem promising. The helical "enzyme" of Chakravarty et al.[38] had *circa* 50% of the activity of hen egg-white lysozyme (w/w), with action on M. lysodeikticus cell walls, chitin, and dextran. Their artificial acetylcholine "receptor" also showed strong binding. Gutte et al.[39] obtained 2.5% of natural RNase activity from the *dimer* of their artificial nuclease. While natural RNase digests polyC, polyU, and polyA, in that order, the artificial enzyme digested polyC, polyA, polyU, and polyG, in that order. There was a preference for cleavage at the C′-end of polyC, and evidence of strong preferential binding of cytidine polyphosphates. The monomeric form of the enzyme was less active and, interestingly, bound 2′-CMP some 30 times less strongly.

However, to assess the contribution made by rational design, one should assess the effects of a random organization of amino acids uninfluenced by design. Randomly polymerized amino acids[45] have both significant hydrolytic activities of various types[46] and, associated with each, a degree of specificity including stereospecificity. Even the simplest attempts at order can produce an improvement; simple glutamate copolymers have some lysozyme-like activity[47,48] and formed the basis of the study of Dhar and colleagues.[38] Comparison with a random sequence containing amino acids in the same proportions as the designed enzyme should provide a suitable baseline.[1] It seems surprising that these random structures are active since they are unlikely to have the specific groups in the right relative positions. Probably, the substrate itself induces the required arrangement. In fact, the most convincing evidence for this is found in studies by the designers themselves.

First, consider the α-helical, artificial acetylcholine receptor of Dhar and colleagues.[43] In accord with our earlier comments on the instability of short helices, the helix content was shown to *increase* by polarimetry on binding acetylcholine. The sequence was, in this case Glu-Ala-Ala-Ala-Ser, chosen such that the terminal group had the required distance for the binding if the scaffold was, indeed, helical. In the case of Ala-Ala-Glu-Ala-Ser, designed on this rational *not* to bind acetylcholine, the helix content *decreased*. This implies that the binding induced the planned bad binding conformation into a good binding conformation, overriding the intentions of the designer. In short, the correct choice of groups seemed far more profitable than the rational choice of their relative position, because of the important contribution of the binding energy to the structure of the overall complex.

This problem was not avoided in the study of Gutte et al.[39] Unexpected binding specificities plus changes in helix and pleated sheet content, as followed by circular dichroism, obliged the authors to consider the possibility that the intended conformation, and that in the complex, were significantly different. In any event, the binding clearly caused very significant conformational changes though the precise extent and nature of these await further study.

VII. CHAIN-STATISTICAL ASPECTS OF SCAFFOLDS

The above considerations suggest a more statistical view of scaffold design. Either as separate chemical species or as groups on a random polypeptide chain, the catalytic and binding groups always have some chance of being at the right relative position for inter-

acting with the substrate. The design process (and probably evolution), thus, involves optimizing the probability of the required distances.

Statistical treatment of distances between points on a random or partially random polypeptide demands that the root mean square distance $\overline{d^2}$ be calculated.[49,50] If each stereochemical unit is represented by a vector connecting successive C^α carbon atoms, then vector algebra[49] gives

$$\overline{d^2} = \sum_i \sum_j \overline{p_{ij}} \qquad (14)$$

where $\overline{p_{ij}}$ is the mean scalar vector product of vectors representing units i and j.

Those assignments of values of $\overline{p_{ij}}$ which are fixed by polypeptide geometry and which are common to most authors (except for differences in choice of polypeptide geometry such as the distance between successive C^α atoms) are listed in Table 1. Authors[49-52] differ, however, in the interpretation and choice of M (see table) and the precise value to be assigned to $\overline{p_{ij}}$ $(0 < p_{ij} < u^2)$. For a homopolypeptide of one type of unit (for example, a glutamic acid unit), a value of M can be calibrated from the experimental characteristic ratio[50] $C_\infty = \overline{d^2}/nu^2$ when the number of units n is very large, assuming that $\overline{p_{ij}} = 0$ for $|i-j| + 1 > M$ and e^2 for $1 < |i-j| \leqslant M$. Here e is the length of the projection of each vector on the helix axis of the most stretched chain attainable in practice[49] (*circa* 3.55 Å); following Equation 14 we obtain for a chain of n units.

$$\begin{aligned} &M > n, \quad && M = n \\ &M < n, \quad && M = u^2(C_\infty - 1 - 2\cos\theta)/(2e^2) + 2 \end{aligned} \qquad (15)$$

The second gives the only value of M satisfying $d^2 = C_\infty nu^2$, for large n. By inspection of summation in Equation 14 for n, in general, we obtain

$$\overline{d^2} = nu^2 + (2n-2)u^2\cos\theta + [2Mn - M(M-1) - 4n + 2]e^2 \qquad (16)$$

This derivation gives a good approximation to the approach of Brant and Flory[50] and Kratky and Porod,[52] and, indeed, is an improvement on the latter for short chains. However, the reason for presenting it here is that it clarifies the treatment of other workers. Brant and Flory,[50] for example, related p_{ij} to $(T^k)_{11}$ where $k = |i-j| + 1$, T is the averaged orthogonal matrix which expresses each vector in the coordinate system of its predecessor, and 11 denotes the first element of the product k-such matrices. The matrix T is averaged by statistical mechanical averaging which requires calculation of the energies of interaction between units. The significance of this is that it does not assume some critical separation M between units along the sequence at which $\overline{p_{ij}}$ will fall from e^2 to zero. In other words, the degree of stiffness of any section is taken as a function of its length k. As applied, it does assume that the net energy of interaction between units is only significant between i and i + 2 (i with i + 1 being determined by geometry), but it is easy to show that this is not a requirement for random coil behavior. For example, Equation 16 still gives the required convergence to $C_\infty nu^2$ for long chains, providing n greatly exceeds M, but M could still be greater than 2 and imply significant net interactions between i and j if not too widely separated. The Brant-Flory approach provides a better account of heteropolypeptides, where a different matrix T must be calculated for each type of residue associated with a unit. To adapt the simple approach based on M one would have to use Equation 14 with $\overline{p_{ij}}$ set explicitly, using zero rather

Table 1
ASSIGNMENT OF VALUES TO THE SCALAR PRODUCTS OF VECTORS REPRESENTING POLYPEPTIDE UNITS

Condition	Value (Å^2)	Notes
$i = j$	u^2	u = 3.8 Å is the length of the vector spanning adjacent C^α atoms; the product of a vector with itself is the square of its magnitude
$\|i - j\| = 1$	$u^2 \cos \theta$	θ is the angle between vectors describing adjacent units, fixed by molecular geometry such that $\cos \theta = 0.48$
$\|i - j\| + 1 > M$	0	M is some critical separation between units in the sequence such that vectors are uncorrelated, i.e., no net interactions and do not belong to same stiff or partially stiff chain section; average value of $\cos \theta$ is zero when vectors then move at random with respect to each other
i and j belong to same stiff stereoregular section	h^2 approximately	To a good approximation, vectors can be replaced by their projections on the helix axis, of length h (equivalent to the rise per residue)

than e^2 if $|i - j| + 1$ exceeds a value of M which is now characteristic of the intervening sequence. One might use, for M,

$$\left\{ \sum_{g=i}^{j} M_g^{-1} \right\}^{-1}$$

where M_g is the characteristic value of M for the appropriate unit in its homopolypeptide, or take M as the lowest vlaue of M_g encountered in the intervening units. The method of choice would depend on the model for the nature of the interactions between units. The Brant-Flory treatment, however, allows no such laxity of choice, though it could be argued that the particular limit set for the range of energy interactions is an arbitrary choice of this type and a much less justified choice in the heteropolypeptide case. The Kratky-Porod model[52] is closer to the simple treatment using M and there is an analogous quantity calibrated against C_∞. The difference is that the chain is smoothly bending rather than composed of individually rigid vectors and M is replaced by the minimal length of chain over which a certain specified degree of bending can occur. This treatment is responsible for the failure of their model at short chain lengths, where the real unit character of the chain dominates the situation.

The use of these ideas in the design context can be discussed in relation to the model nuclease designed by Gutte et al.[39] Here, certain sections of sequence were designed to be stereoregular by choosing amino acid residues with a strong propensity for particular types of stereoregularity. Residues 1 to 7 and 10 to 16 were extended chain formers; 21 to 34 were helix formers. In this case, the remaining residues were selected as reverse turn formers and hydrophobic residues were judiciously sited in the stereoregular regions in order to try an induce association of the stereoregular sections. This would have the effect of introducing $\overline{p_{ij}}$ values greater than zero between units far apart in the sequence, but here we ignore this both for simplicity and because the actual result may, indeed, have been more flexible than intended, as discussed above. Instead, the intervening

regions 8 and 9 and 17 to 21 will be considered as "universal joints" implying $\overline{p_{ij}} = 0$ when either/or both lie between i and j. Then, following Equation 16 and treating stereoregular sections as single long units of length L = nh (where n units in each stereoregular structure have a projection length h on the structure axis), we obtain

$$\overline{d_2} = L_A^2 + L_B^2 + L_H^2 + n_t u^2 + (2n_t - 2)u^2 \cos\theta \qquad (17)$$

Here A,B,C relate to the two extended chains and helix, respectively, and n_t is the total number of "turn" residues not in stereoregular structures. Mean square end-to-end distances of any part of the structure are obtained in the same way, considering the stereoregular structures and n_t in that chain section.

There are two problems here. The first is that we cannot guarantee that short stereoregular structures such as α-helix will be stable. To take account of this one has to establish weighting coefficients a,b,c (lying between zero and unity) and the end-to-end distances of the random coil sections R with which each stereoregular section is in equilibrium:

$$\overline{d^2} = aL_A^2 + (1-a)R_A^2 + bL_B^2 + (1-b)R_B^2 + cL_H^2 + (1-c)R_H^2 + n_t u^2 + (2n_t - 2)u^2 \cos\theta \qquad (18)$$

This problem, though theoretically tractable[49] using theories like those of Zimm and Bragg[53] is difficult and seems to justify use of random chains without stereoregular sections, for simplicity. It would, however, be possible to use the above equation if the end-to-end distances of constituent parts were obtained experimentally by hydrodynamic or light-scattering techniques, the required unknowns determined empirically from this data and assumed valid for the complete structure. Unfortunately, helix-coil transition theory which might lead theoretically to a,b,c includes use of a cooperativity parameter which is very sensitive to interhelical interactions,[54] and one can envisage the same problem for extended chain sections. Another problem is that it would seem difficult to guarantee "universal joint" behavior for the residues not in stereoregular sections. One might expect any real residue to introduce stiffness into the joints which would cause the units in separate stereoregular sections to become correlated, invalidating the above equations.

However, there are a number of studies, notably those of Hagler and Honig[55] and Hagler et al.,[56] which emphasize that alanine and alanine-like residues tend to form rather stiff near-extended chains into which universal joints can be introduced by glycine. This behavior of glycine is consistent with C_∞ of *circa* 2 for polyglycine,[49] implying an M = 2 in Equation 16 and effectively zero p_{ij} except where determined by local geometry. By comparison of calculated and observed behavior of polypeptides with guest glycine residues,[56] the value of these ideas in design has been demonstrated. Further, Hagler and Honig[55] have also discussed the important role of glycine as a "universal joint" in the evolutionary "design" process. We may conclude that for short polypeptides, at least where interactions between residues far apart in the sequence do not contribute significantly, reasonable estimates of statistical behavior can be expected when glycine is introduced at specific sites. Further, when such interactions do occur, glycine still exhibits hinging behavior and alanine residues still form fairly stiff chains such that the number of conformational possibilities is greatly restricted *a priori*.[55] Indeed, it is reasonable to suppose that Gutte et al.[39] would have obtained similar results simply using glycine for every intended turn residue.

VIII. CALCULATION OF SECONDARY STRUCTURE

Secondary structure prediction methods have been reviewed extensively in the literature.[57] They include those based on helix-coil transition theory and at least some experimentally determined parameters,[58,59] those based on stereochemical rules,[60–62] and those based on parameters derived from analysis of proteins of known sequence and conformation.[63–71] The latter set is particularly popular. It may be further loosely divided into those which are unambiguous, easily programmed, and based on rigorous statistical reasoning, but difficult to do without at least a microcomputer, and those which are easy to carry out manually but with some small variation in results from user to user, since rules have been expressed in language with some flexibility in precise interpretation. The latter are particularly useful in design work where one is choosing the sequence to give a particular secondary structure, since the rules do give clear instruction about which kind of sequences should produce a desired effect, even if their qualitative nature provides difficulties in a few marginal cases when used in the classic prediction mode. Gutte *et al.*[39] used the popular Chou-Fasman[70] approach.

The heart of the popular statistical approach is the calculation of parameters from the data base of proteins of known sequence and conformation. These parameters represent the propensity,[68,70] or the information,[63–67] for a particular type of residue (for example, alanine) having a particular type of conformation (for example, α-helical). In the long-established information approach such a parameter would be given by[65–67]

$$I(\text{helical; alanine}) = \#(n_{ha}) - \#(n_{\bar{h}a}) - \#(e_{ha}) + \#(e_{\bar{h}a}) \qquad (19)$$

where n_{ha} and $n_{\bar{h}a}$ are the numbers of helical and nonhelical alanine residues, respectively, and e the corresponding "expected" numbers in the sense used in the chi-square test, i.e., e_{ha} = number of helical residues × number of alanine residues/total number of residues. The function # is defined as

$$\begin{aligned} &\#(0) = 0 \\ &\#(1) = 1 \\ &\#(n) = 1 + 1/2 + 1/3 + \cdots \frac{1}{n} \end{aligned} \qquad (20)$$

Which takes account of statistical significance.[65] In the prediction mode, for each residue in the new sequence one writes down the corresponding parameter. What happens next depends on the method, but in a simple approach a residue might be assigned as helical if it belonged to a run of, for example, four residues for which the sum of the parameters is greater than zero. Differences in method relate to cooperativity between residue conformers,[53,54,72] special effects between specific residues,[64,66] and with what to do when one is interested in more than one conformational possibility (i.e., if assignments are not simply helical or nonhelical).[67] In the design procedure, one simply writes the residue which will have the highest propensity for the required conformation, though subject to other chemical requirements. However, because of possible effects due to cooperativity, interaction between specific residues, and interference between conformational possibilities (i.e., on equilibria between different conformations), one must subsequently still carry out a prediction to check that the result is consistent with requirements. Ideally, an interactive prediction program should be used to explore the effects of sequence modifications.

The nature of the parameters, nonetheless, poses a limitation on the value of predictions. First, unless interactions between many residues are considered, i.e., unless

tertiary structure effects are included, the results are unlikely to do much better than assigning 75 to 80% of residues to the correct conformation.[73] Second, the parameters describe the propensity, not to a specific backbone geometry, but to a broad conformational state which is the range of geometric variables which encompasses the observed variation in that type of structure. Even in the case of a perfect prediction when each residue is assigned to its observed state, there is insufficient data to build up even an approximate representation of the tertiary structure.[73] In principle, both these deficiencies could be overcome with the type of parameter one might obtain from a very large data base, but in any future we can foresee that the predictions can only re really useful as starting points for more detailed consideration, preferably a calculation of the structure by an energy minimization procedure (see below). Nonetheless, it should be noted that useful information can be obtained by statistical analysis of interatomic or intersidechain distances,[74,75] which include some tertiary effects.

IX. ENERGY CALCULATIONS AND CALCULATION OF TERTIARY STRUCTURE

Most of the theoretical principles determining molecular structure and behavior have long been understood as discussed in 1929 by Dirac.[76] In principle, we have all the equations and all the data we need for calculating secondary and tertiary structure of scaffolds, the free energy of association of molecular species, and catalytic processes and specificities, providing the covalent structures are known. In practice, to carry out calculations (solve these equations) from first principles is impossible because of the enormous amount of computer time required. Thus, many approximations are made, many contributions neglected, and much emphasis laid on the use of empirical data to by-pass time-consuming steps in calculation. The kind of use of empirical data as envisaged in the prediction of secondary structure above is, nonetheless, severely limited by the number of experimental observations that would be required to treat more complex problems. In particular, it would be unfeasible to calculate tertiary structures of scaffolds purely on statistical grounds, since the data required to account for all combinations of group positions in space would be enormous. Thus, it is more profitable to go back a little further in the direction of using fundamental principles, sacrificing computer time to compensate for the lack of information in directly relevant empirical observations.

The type of calculation we envisage is an energy calculation, since free energies of equilibrium systems and the time-course of events in nonequilibrium systems can all subsequently be calculated from the potential energies of the system.[25-27] In Table 2 are listed some examples of energy calculations used to treat specific problems concerning peptide and polypeptide structures. These were not primarily design studies, but they could have formed part of a design study and the results obtained could be of value in future design studies. They are arranged in order of increasing structural complexity, from low molecular weight peptides to proteins.

More profitable for present purposes, however, is a brief account of different approaches in order of the nature and severity of the approximations made. Since the least approximate are both more expensive and less routine, we can deal with and eliminate these first.

Ideally, all energies E of any state of the system can be calculated from the eigenfunction-eigenvalue equation:

$$H\psi = E\psi \tag{21}$$

where ψ is the wave function, E the energy, and H is the Hamiltonian operator. The

Table 2
SOME RECENT ENERGY CALCULATIONS WHICH ARE RELEVANT TO DESIGN PROBLEMS

System studies	Notes	Ref.
Peptide group	Compare calculated and crystal structures of $CH_3.CO.NH.CH_3$; Zimmerman-Scheraga consider peptide group cis-trans conversions in dipeptides	77—79
Blocked residues, (i.e., N-acetyl amino acyl N′-methylamides)	The first three of these use empirical potential functions, the next two use *ab initio* methods; the last study is directed towards calculation of some experimentally determined properties (e.g., NMR data)	80—84
Blocked dipeptides, (i.e., N-acetyl amino acyl N′-methylamides) and tri- and tetrapeptides	These studies were primarily aimed at studying the relative importance of local and tertiary structure interactions on chain inversions (β-bends) in globular proteins	81, 85—87
Melanotropin H-Pro-L-Leu-Gly-NH_2	These workers calculated a conformation consistent with X-ray data and NMR; β-bend type	88
Thyrotropin Pyroglu-His-Pro-NH_2	Agreement with IR, NMR; again β-bend type	89
Enkephalin Tyr-Gly-Gly-Phe-Met and Tyr-Gly-Gly-Phe-Leu	The last three studies rely on pharmacological data or reasoning, e.g., binding/activity data, or knowledge that the relatively rigid morphine molecule has a similar action	4, 90—93
Lutinizing-hormone releasing factor (decapeptide)	Illustrates general attack for larger drug peptides; calculations are first done on fragments, which define suitable starting conformations for minimization	94
Cyclic peptides	Much work has been done in this area; ring closure reduces the number of possible answers, but provides technical difficulties	95, 96
Collagen	Recent work on fibrous proteins has been largely directed to collagen because of the greater challenge posed by poly(Gly-Pro-Pro)	97, 98
X-ray refinement of globular proteins	This procedure is now applied almost routinely, though result depends on potential functions used and need not always imply better fit of results to crystal data	99—101
Protein vibrations	This study on the small protein pancreatic trypsin inhibitor used molecular dynamics methods to study vibrational motion in the native state; technique generally limited to less than 10^{-10}-sec timescale, and not yet suitable for structure prediction when many atoms are involved	102
Prediction of protein structure from that of homolgous proteins	Predict α-lactalbumin from lysozyme, neurotoxic protease inhibitors from pancreatic trypsin inhibitor, and thrombin and trypsin from elastase, respectively; last study[105] could have used homologous myoglobins equally well.	103—105
Prediction of protein structure from amino acid sequence	Although the aim is to use amino acid sequence alone, many of these studies did exploit some prior knowledge of the result; in all cases, agreement with the observed structure is very crude	106, 107 116, 117
Water-peptide/protein interactions	These studies all used the Monte Carlo technique; the first two employed hydrated crystals, the last a peptide solution	108, 114, 115

latter is the input to the calculation of E. It represents a recipe for the determination of E and is obtained (1) by writing the classical (Newtonian) expression for the energy of the system, (2) by rewriting to make explicit the dependance of the energy on the momenta

of the particles in the system, and (3) by replacing each momentum by the quantum mechanical momentum

$$\frac{-ih}{2\pi} \cdot \frac{d}{d_q}$$

where d/dq is the instruction to differentiate the wave function with respect to particle coordinates q.

For a many-particle system, the closest approach to this level of exactness which can be used in practice is an *ab initio* calculation. This is sufficiently time consuming to allow only the estimation of the energy of one conformation (or very few) of one small peptide-type molecule at a time (i.e., in about 1 hr), and one could not possibly search out the most stable structure of a polypeptide scaffold by this approach. As it is, many approximations are still made. Briefly, one starts with the atomic orbitals, each represented by a number of Gaussian or Slater functions added together, and, using an approach which implies exploitation of a perturbation approximation for many-electron interactions, one seeks out the electronic configuration of least energy. Calculations using extended basis sets (i.e., with extra functions to give a very flexible description) have been used to calculate potential energy surfaces (as a function of conformation) of glycine and alanine dipeptide analogs.[82,83]

The reason for such studies is to obtain more approximate, rapid methods which can handle a much larger number of conformations in reasonable time. Specifically, one seeks *potential functions*[106,107] which are analytical representations of the energy of interaction between atoms, typically dependent on atom type and the distance between the atom centers. These can be obtained from experimental data[109,110] but must be shown to give results suitable for transfer to conformational calculation, e.g., by comparison with experiment and *ab initio* results.[83,84] Although experimentalists regard this use of the *ab initio* approach with some concern, it enjoys special status because of its dependence on fundamental principles and the absence of any *ad hoc* input parameters. Though the functions modeling atomic orbitals are, in a sense, input parameters, they actually represent an arbitrary starting point for calculation of the electronic configuration of least energy, and, ideally, should not affect the result. In practice, of course, they do, since the degree of flexibility of the functions and the time available for varying them is finite. Their importance is that they provide additional data, and in some cases the only available data, for development of the potential functions.

Given a good set of potential functions, one can calculate the potential energy of any conformer as the sum of its pairwise interatomic interactions, possibly combined with analytical functions for representation of bond-stretching, valence angle bending.[84] The latter are also readily obtained from *ab initio* calculation,[83] though here infrared spectroscopy provides a convenient, experimental alternative.[84] One is forced, however, to neglect quantization; the potential functions give the energy as continuous. In this sense the potential functions assume the classical (Newtonian) behavior of the system when the Hamiltonian becomes simply the total (potential plus kinetic) energy of the system. More cautiously, one should say that quantization could be introduced retrospectively, e.g., to determine behavior in a potential well,[56] but the important point is that at least part of the potential energy surface must be determined, first assuming classical behavior. The most fundamental calculation which neglects quantization is Molecular Dynamics,[102] which has the considerable merit of including the kinetic energy of the nuclei and, in this one respect, has advantages even over the *ab initio* calculation as normally applied. Molecular Dynamics works by moving the atoms according to algorithms which imply Newton's laws, given a suitable starting point, a specified time step, and a temperature which is calibrated by scaling the atomic velocities. The inclusion of kinetic energy allows

a more realistic estimate of free energies of stable and metastable conformations as well as giving a description of the time course of events (including the frequencies of conformational vibrations). To treat processes lasting 10^{-10} sec of real time may take many hours of computer time, so the apparent advantage of allowing a study of nonequilibrium processes (such as the folding of a protein towards its equilibrium condition) is somewhat illusionary at present.

Monte Carlo is somewhat cheaper and neglects kinetic energy. Since atoms have no momenta,[111-113] one may sample as one wishes and the name "Monte Carlo" implies that this is effectively done at random to avoid any bias introduced by prior prejudice. More precisely, one puts in a bias and extracts its effects later, since random sampling would be very time consuming. Sampling is completed when the properties of interest have converged within a present criterion of accuracy. The methods available differ according to the method of biasing and extraction, and include biasing according to the known behavior of the units of the polypeptide[113] and the Metropolis[111] algorithm. Like Molecular Dynamics, it is an excellent method for considering the average properties of a very flexible system, such as random scaffolds[112-113] and the solvent component of peptide solutions.[108,114-115]

If one is interested only in the most stable conformation, then minimization of the energy as a function of conformation will, in principle, suffice.[55,116-118] Although this procedure locates minima in the potential energy surface rather than calculates free energies, the latter can be calculated at the minima once found.[56] This seems the method of choice for calculating the expected structure of a rigid polypeptide scaffold. Although energy minimization is the most economic and feasible approach, it is still limited by the complexity of the conformational energy surface of a protein, i.e., by the large number of local minima in the energy as function of the conformational variables. Classical gradient minimization methods will only take the scaffold to the nearest minimum, so there must be facilities for escape and searching for the deepest minimum.[116,117] The SIMPLEX minimization method will accelerate the escape from minima in a natural way,[117] and though normally rather slow, it can be combined with classic gradient methods which locate the nearest minimum very rapidly.[117]

The complexity of the energy surface, the need for good potential functions, and the need to explicitly represent water molecules[84] has meant that no successful prediction of a protein's tertiary structure has been made from amino acid sequence alone, using minimization or any other method. This does not bode well for the use of this approach to scaffold design, though it should be recalled that the induced fit of artificial enzyme by substrate can make up for many deficiencies. The point remains, nonetheless, as to whether such detailed calculation is any better than subjective judgement, using empirical rules of thumb. In the current status of the art, the answer is probably no. However, a less detailed but automatic calculation may be an improvement and would benefit from being objective and reproducible, giving the same results for the same computer program in the hands of different authors. For example, it would allow better judgment of whether two secondary structure features had surface residues in the correct positions for a favorable interaction. In complex cases involving, for example, two helices and a pleated sheet, it might be rather difficult to judge the correct spatial arrangement of the interacting groups. Thus, an interactive program including visual display could be a tremendous asset to the designer.

X. INCLUSION OF SOLVENT EFFECTS IN THE CALCULATION OF TERTIARY STRUCTURE

Since even the fastest technique for calculating the most stable conformation of a scaffold, namely, energy minimization, has not been successful as yet for complex

polypeptide systems, it would seem ambitious to take detailed account of the role of the water solvent. Each water molecule would introduce new degrees of freedom, and many hundreds of water molecules may need to be considered. Further, it is not particularly useful to minimize the energy of the solvent system as a function of conformation, since, except in ice, there is no one overall conformation favored above any others. One would have to apply a solution Monte Carlo simulation[108,114-115] or the even more expensive Molecular Dynamics technique[119] for *each* solute conformer encountered in the course of minimization of the polypeptide solute. No such study for many solute conformers has yet been attempted, and to do this combined minimization and Monte Carlo approach properly and efficiently would almost certainly not be trivial.

Fortunately, some important effects of the solvent can be included within the intramolecular potential functions and explicit representation of water can be neglected, at least as a first approximation. Traditionally, many workers have replaced or extended the van der Waals' interactions between nonpolar atoms or groups to represent hydrophobic interactions (e.g., Levitt and Warshel,[116] Robson and Osguthorpe,[117] and Nemethy and Scheraga[118]), while in a great number of studies electrostatic interactions between charged or partially charged atoms or groups have been reduced by introduction of a dielectric constant (e.g., Brant and Flory[50]). The problem is that many such treatments are both quantitatively and qualitatively poorly justified; more exact calculations on peptide-solvent systems are required both to justify or refute current approaches and to suggest improvements. To this end, Monte Carlo studies were carried out both on hydrated polypeptide crystals[108,114] and on a dipeptide in solution.[115] Periodic boundary conditions, in which the solution volume is represented as an array of unit cells of identical conformation, were used to model infinite solution phases and to avoid solution-vacuum interfaces.

The first conclusion is that in case of a dipeptide in solution, the solvent behaves very much like pure water except in the innermost shell around the solute (i.e., within about 3.2Å of the solute surface).[115] This is not true for hydrated protein crystals, where "glue channels" of ordered water[114] play the major role in stabilizing the orientation and distance between adjacent protein molecules, and, thus, in stabilizing the crystal structure as a whole. However, it suggests that *solvation shell* models,[120,121] in which the solvent-dependent contribution may be made a function of the volume of intersection of two hydration shells[116,117] as the groups supporting them approach, may be reasonable. The idea is that the special solvation shell water is displaced from between the groups and is restored to the bulk solvent with a change in its thermodynamic properties. If the free energy change for removing all the solvation shell water is deduced from experimental studies of amino acid or peptide solubility in aqueous and nonaqueous media, then the free energy for removing part of it is easily calculable and held to represent the strength of the group interaction.[116,117]

The second conclusion is that the *supermolecule approximation*[122,123] may also be reasonably well founded. This approximation treats any solute and strongly associated water molecules as a single, giant molecular species. The Monte Carlo simulations reveal that water molecules are strongly bonded to peptide hydrogen bonding groups. The problem is that the link is still much weaker and more easily deformed than covalent bond, so that this inherent flexibility must be taken account of if attached water molecules are to be treated as part of the scaffold. Preferably, one should also allow water molecules to be removed completely when displaced by steric contacts with other groups in the polypeptide.

The third conclusion is that the *Onsanger reaction field*[33,124-126] can be of immense, even dominating, importance. This arises from the weak interaction of solute with each water molecule a considerable distance away, but it is a very significant overall contribution for the progressively larger number of water molecules encountered as one works out-

ward through the concentric solvation shells around the solute. The net effect is that a polar solvent like water will favor solute conformers with the largest dipole moments, other factors being equal. Fortunately, being an effect due to many molecules at a distance, it is statistical in nature, allows the solvent to be considered as a continuum, and can be treated as a simple mathematical function of the dielectric constant of the medium and the dimensions of the cavity in the solvent which the solute occupies.

The change in effective cavity size of a smaller ligand as it enters the cavity of the larger ligand may be sufficient to alter the conformation of the smaller ligand to one with a smaller intrinsic dipole moment, and this should be considered in assessing the overall free energy of complex formation.

XI. VIBRATIONAL FREE ENERGY

Another important factor which must be considered is the free energy contribution due to the vibrational entropy polypeptide itself. The changes in free energy resulting from this contribution can be calculated and shown to be of the same order as the enthalpic contribution.[56] A chain in which n monomers each have m equivalent potential energy minima would, in fact, undergo a free energy increase nRTln(m) in folding to a form in which vibration is confined to just one of these minima per monomer, so for a 25-residue polypeptide a free energy increase of the order of 100 kJ (mole scaffold)$^{-1}$ would not be unexpected. It is this high free energy which must be outweighed by the intramolecular potentials used and solvent effects.

XII. CONCLUSIONS

In this review I have examined some general design principles, some current design attempts, and some more exact methods of calculation which can be applied to the design problem. On the one hand, it might be felt that some significant successes have been achieved in design without recourse to detailed calculation, and that since the important thing is the quality of the result, then more exact methods are superfluous. On the other hand, these attempts have been few, and I have argued that enzyme design attempts may have been less significant than first appears because of the natural tendancy for a substrate to induce the fit of the artificial enzyme. These deficiencies may be *because* detailed calculation has not been carried out.

Although design is primarily a problem in theoretical chemistry because the molecule does not yet exist, it cannot remain a theoretical problem because the aim is to make the molecule exist. So far, the real design attempts have been carried out by experimental groups. However, both design and synthesis are time consuming and require special expertise. There can be no one designer; rather, it is a task for an integrated group. Such integrated groups, as there are so far, belong to the pharmaceutical industry, and it seems timely to consider a more integrated theoretical and practical approach in other areas.

REFERENCES

1. **Robson, B.**, *Trends Biochem. Sci.*, 5, 240, 1980.
2. **Robson, B.**, *Trends Biochem. Sci.*, 3, 49, 1976.
3. **Sarett, L. H.**, *Res. Manage.*, p. 18, 1974.
4. **Beddell, C. R., Clark, R. B., Lowe, L. A., and Wilkinson, S.**, *Br. J. Pharm.*, 61, 351, 1977.
5. **Monahan, M. W., Amoss, M. S., Anderson, H. A., and Vale, W.**, *Biochemistry*, 12, 4616, 1973.
6. **Redell, G., Cramer, R. D., and Berkoff, C. E.**, *Chem. Soc. Rev.*, p. 273, 1974.
7. **Young, P. A.**, Communication to the biometric society (British Region), discussed in *Br. J. Pharm.*, 61, 351, 1976.
8. **Chaplin, M. F.**, *Trends Biochem. Sci.*, 6(1), 1981.

9. **Robson, B.,** *Trends Biochem. Sci.*, 6(5), 1981.
10. **Chaplin, M. F.,** *Trends Biochem. Sci.*, 6(5), 1981.
11. **Lipscomb, W. N.,** Quantum biochemistry in biomedical sciences, *Ann. N. Y. Acad. Sci.*, 367, 1981.
12. **Tolkovsky, A. M. and Levitski, A.,** *Biochemistry*, 17, 3795, 1978.
13. **Levitski, A., and Helmreich, E. J. M.,** *F.E.B.S. Lett.*, 101, 213, 1979.
14. **Jacobs, S. and Cuatrecassas, P.,** *Biochem. Biophys. Acta*, 433, 482, 1976.
15. **Snedecar, G. W. and Cochran, W. G.,** *Statistical Methods*, Iowa State University Press, 1967.
16. **Anderson, T. W.,** *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, New York, 1958.
17. **Nilsson, N. J.,** *Learning Machines*, McGraw-Hill, New York, 1965.
18. **Kier, L. B.,** *Advances in Chemistry Series*, Vol. 14, American Chemical Society, Washington, D.C., 1972, 319.
19. **Montgomery, J. A., Mayo, J. G., and Hansch, C.,** *J. Med. Chem.*, 17, 477, 1974.
20. **Free, S. M. and Wilson, J. W.,** *J. Med. Chem.*, 7, 395, 1964.
21. **Marshall, G. R. and Bosshard, H. E.,** *Circ. Res.*, 30 and 31 (Suppl. 2), 143, 1972.
22. **Venkatachalam, C. M.,** *Biopolymers*, 6, 1425, 1968.
23. **Bradbury, A. F., Smyth, D. G., and Snell, C. R.,** *Nature*, 260, 165, 1976.
24. **De Coen, J. L., Humblet, C., and Koch, M. H.,** *F.E.B.S. Lett.*, 73, 1977.
25. **Hill, T. L.,** *An Introduction to Statistical Mechanics*, Addison-Wesley, London, 1960.
26. **Gibbs, J. W.,** *The Collected Works of J. Willard Gibbs*, Green, London, 1931, 3.
27. **Rushbrooke, G. S.,** *Introduction to Statistical Mechanics*, Clarendon Press, Oxford, 1949.
28. **Burgen, A. S. U., Roberts, G. C. K., Feeney, J.,** *Nature*, 253, 753, 1975.
29. **Weinstein, H., Maayani, S., Cohen, S., and Sokolovstry, M.,** *Mol. Pharmacol.*, 9, 820, 1973.
30. **Burt, S. K., Loew, G. H., and Hashimoto, G. M.,** Quantum biochemistry in biomedical sciences, *Ann. N.Y. Acad. Sci.*, 367, 1981.
31. **Weinstein, H., Osman, R., Topiol, S., and Green, J. P.,** Quantum biochemistry in biomedical sciences, *Ann. N.Y. Acad. Sci.*, 367, 434, 1981.
32. **Robson, B., Douglas, G., Metcalfe, A., Woolley, K., and Thompson, J. S.,** Proc. Symp. Biophysics of Water, Cambridge, 1981. in *Biophysics of Water*, Franks, F., Ed., John Wiley & Sons, in press, 1982.
33. **Onsanger, L.,** *J. Am. Chem. Soc.*, 58, 1486, 1936.
34. **Crippen, G. M.,** *J. Mol. Chem.*, 22, 988, 1979.
35. **Blundell, T. L., Dodson, G. G., Hodgkin, D. C., and Mercola, D. A.,** *Adv. Protein Chem.*, 26, 279, 1972.
36. **Sasaki, K. S., Dockerill, S., Adamiak, D. A., Tickle, I. J., and Blundell, T.,** *Nature*, 257, 751, 1975.
37. **Boyd, D. B.,** Quantum biochemistry in biomedical science, *Ann. N.Y. Acad. Sci.*, 367, 531, 1981.
38. **Chakravarty, P. K., Mathur, K. B., and Dhar, M. M.,** *Indian J. Chem.*, 12, 464, 1973.
39. **Gutte, B., Däumigen, M., and Wittschieber, E.,** *Nature*, 281, 650, 1979.
40. **Warshel, A. and Levitt, M.,** *J. Mol. Biol.*, 103, 227, 1976.
41. **Warshel, A. and Weiss, R. M.,** Quantum biochemistry in biomedical sciences, *Ann. N.Y. Acad. Sci.*, 367, 370, 1981.
42. **Allen, L. C.,** Quantum biochemistry in biomedical sciences, *Ann. N.Y. Acad. Sci.*, 367, 383, 1981.
43. **Chakravarty, P. K., Mathur, K. B., and Dhar, M. M.,** *Indian J. Biochem. Biophys.*, 10, 233, 1973.
44. **Bayer, E. and Holzbach, G.,** *Angew. Chem.*, 89, 120, 1977.
45. **Fox, S. W. and Harada, K.,** *Science*, 128, 1214, 1958.
46. **Fox, S. W. and Rohlfing, D. L.,** *Adv. Catal.*, 20, 373, 1969.
47. **Dhar, M. M., Agrawal, A. K., Mathur, K. B., Ray, C.,** *Indian J. Biochem. Biophys.*, 10, 227, 1973.
48. **Chakravarty, P. K., Mathur, K. B., and Dhar, M. M.,** *Experentia*, 29, 786, 1973.
49. **Flory, P. J.,** *Statistical Mechanics of Chain Molecules*, Interscience, New York, 1969.
50. **Brant, D. A. and Flory, P. J.,** *J. Am. Chem. Soc.*, 87, 2791, 1965.
51. **Birshtein, T. M. and Ptitsyn, O. B.,** *Conformations of Macromolecules*, Timashaff, S. N. and Timashaff, M. J., Eds., Interscience, New York, 1966.
52. **Kratky, O. and Porod, G.,** *Rec. Trav. Chim.*, 68, 1106, 1949.
53. **Zimm, B. H. and Bragg, J. K.,** *J. Chem. Phys.*, 31, 526, 1959.
54. **Suzuki, E. and Robson, B.,** *J. Mol. Biol.*, 107, 357, 1977.
55. **Hagler, A. T. and Honig, B.,** *Proc. Natl. Acad. Sci.*, 75, 554, 1978.
56. **Hagler, A. T., Stern, P. S., Sharon, R., Becker, J. M., and Naider, F.,** *J. Am. Chem. Soc.*, 101, 6482, 1979.
57. **Schultz, G. E., and Schirmer, R. H.,** *Principles of Protein Structure*, Springer-Verlag, Basel, 1979.
58. **Lewis, P. N., Gō, N., Gō, M., Kotelchuck, D., and Scherage, H. A.,** *Proc. Natl. Acad. Sci.*, 65, 810, 1970.
59. **Finkelstein, A. V., Ptitsyn, O. B., and Kozitsyn, S. A.,** *Biopolymers*, 16, 497, 1977.

60. **Schiffer, M. and Edmundson, A. B.,** *Biophys. J.*, 7, 121, 1967.
61. **Lim, V. I.,** *J. Mol. Biol.*, 88, 857, 1974.
62. **Lim, V. I.,** *J. Mol. Biol.*, 88, 873, 1974.
63. **Pain, R. H. and Robson, B.,** *Nature*, 227, 62, 1970.
64. **Robson, B. and Pain, R. H.,** *J. Mol. Biol.*, 58, 237, 1971.
65. **Robson, B.,** *Biochem. J.*, 141, 853, 1974.
66. **Robson, B. and Suzuki, E.,** *J. Mol. Biol.*, 107, 327, 1977.
67. **Garnier, J., Osguthorpe, D. J., Robson, B.,** *J. Mol. Biol.*, 190, 97, 1978.
68. **Nagano, K.,** *J. Mol. Biol.*, 75, 401, 1973.
69. **Kabat, E. A. and Wu, T. T.,** *Proc. Natl. Acad. Sci.*, 70, 1473, 1973.
70. **Chou, P. Y. and Fasman, G. D.,** *Biochemistry*, 13, 222, 1974.
71. **Argos, P., Schwarz, J., and Schwarz, J.,** *Biochim. Biophys. Acta*, 439, 261, 1976.
72. **Lifson, S. and Roig, A.,** *J. Chem. Phys.*, 84, 1963, 1961.
73. **Burgess, A. W. and Scheraga, H. A.,** *Proc. Natl. Acad. Sci.*, 72, 1221, 1975.
74. **Warme, P. K. and Morgan, R. S.,** *J. Mol. Biol.*, 118, 273, 1978.
75. **Crampin, J., Nicholson, B. H., Robson, B.,** *Nature, 272, 558, 1978.*
76. **Dirac, D. A. M.,** *Proc. R. Acad. Sci. Ser. A*, 123, 714, 1929. "The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble".
77. **Hagler, A. T., Leiserowitz, L., and Tuval, M.,** *J. Am. Chem. Soc.*, 98, 4600, 1976.
78. **Zimmerman, S. S. and Scheraga, H. A.,** *Macromolecules*, 9, 408, 1976a; *Biopolymers*, 16, 811, 1976b.
79. **Rammachandran, G. N. and Mitra, A. K.,** *J. Mol. Biol.*, 107, 85, 1976.
80. **Lewis, P. N., Momany, F. A., and Scherage, H. A.,** *Biochim. Biophys. Acta*, 303, 211, 1973; *Isr. J. Chem.*, 11, 121, 1973.
81. **Nishikawa, K., Momany, F. A., and Scherage, H. A.,** *Macromolecules*, 7, 797, 1974.
82. **Robson, B., Hillier, I. H., Guest, M.,** *Chem. Soc. Faraday. Trans. II*, 74, 1311, 1978.
83. **Hillier, I. H., and Robson, B.,** *J. Theor. Biol.*, 76, 83, 1979.
84. **Robson, B., Stern, P. S., Hillier, I. H., Osguthorpe, D. J., and Hagler, A. T.,** *J. Chim. Phys.*, 76, 831, 1979.
85. **Melnilzov, P. N., Akhmedou, N. A., Lipkind, G. M., and Popou, E. N.,** *Biorg. Khim.*, 2, 28, 1976.
86. **Howard, J. C., Ali, A., Scherage, H. A., and Momany, F. A.,** *Macromolecules*, 8, 607, 1975.
87. **Hurwitz, F. L. and Hopfinger, A. J.,** *Int. J. Peptide Protein Res.*, 8, 543, 1976.
88. **Ralston, E., De Coen, J. L., and Walker, R.,** *Proc. Natl. Acad. Sci.*, 71, 1142, 1974.
89. **Burgess, A. W., Momany, F. A., and Scherage, H. A.,** *Biopolymers*, 14, 2645, 1975.
90. **Isogai, Y., Nemethy, G., and Scherage, H. A.,** *Proc. Natl. Acad. Sci.*, 74, 414, 1977.
91. **Momany, F. A.,** *Biochem. Biophys. Res. Commun.*, 75, 1098, 1977.
92. **Smith, G. D. and Griffen, J. F.,** *Science*, 199, 1214, 1978.
93. **Gorin, F. A., Balasubramanian, T. M., Barry, D. C., and Marshall, G. R.,** *J. Supramol. Struct.*, 9, 27, 1978.
94. **Momany, F. A.,** *J. Am. Chem. Soc.*, 98, 2990, 1976.
95. **Dygert, M., Gō, N., and Scherage, H. A.,** *Macromolecules*, 8, 750, 1975.
96. **White, D. N. J. and Morrow, C.,** *Comput. Chem.*, 3, 33, 1976.
97. **Miller, M. H. and Scherage, H. A.,** *J. Polym. Sci. (Polym. Symp.)*, 54, 171, 1976.
98. **Okuyama, K., Tanaka, N., Ashida, T., and Kakudo, M.,** *Bull. Chem. Soc. Jpn.*, 49, 1805, 1976.
99. **Levitt, M. and Lifson, S.,** *J. Mol. Biol.*, 46, 269, 1969.
100. **Warme, P. K. and Scherage, H. A.,** *Biochemistry*, 13, 757, 1974.
101. **Swenson, M. K., Burgess, A. W., and Scheraga, H. A.,** *Proc. Symp. Frontiers in Physico-Chemical Biology (Paris)*, 1977.
102. **McCammon, J. A., Gelin, B. R., and Karplus, M.,** *Nature*, 267, 585, 1977.
103. **Warme, P. K., Momany, F. A., Rumball, S. V., Tuttle, R. W., and Scheraga, H. A.,** *Biochemistry*, 13, 768, 1974.
104. **Robson, B. and Timms, D.,** Trends Biochem., Sci., 2, 240, 1980.
105. **Ptitsyn, O. B. and Rashin, A. A.,** *Biophys. Chem.*, 3, 1, 1975.
106. **Tanaka, S. and Scheraga, H. A.,** *Proc. Natl. Acad. Sci.*, 72, 3802, 1975.
107. **Kuntz, I. D., Crippen, G. M., Kollman, O. A., Kimmelman, D.,** *J. Mol. Biol.*, 106, 983, 1976.
108. **Hagler, A. T., Moult, J., and Osguthorpe, D. J.,** *Biopolymers*, 19, 396, 1980.
109. **Hagler, A. T., Huler, E., and Lifson, S.,** *J. Am. Chem. Soc.*, 96, 5319, 1974.
110. **Hagler, A. T. and Lifson, S.,** *J. Am. Chem. Soc.*, 96, 5327, 1974.
111. **Metropolis, N. A., Rosenbluth, A. W., Teller, M. N., and Teller, E.,** *J. Chem. Phys.*, 21, 1087, 1953.
112. **Premilat, S. and Maigret, B.,** *C. R. Acad. Sci. Paris Ser. C*, 282, 225, 1976.

113. **Premilat, S. and Hermans, J.,** *J. Chem. Phys.*, 59, 2602, 1973.
114. **Hagler, A. T. and Moult, J.,** *Nature*, 272, 222, 1978.
115. **Hagler, A. T., Osguthorpe, D. J., and Robson, B.,** *Science*, 208, 599, 1980.
116. **Levitt, M. and Warshel, A.,** *Nature*, 253, 694, 1975.
117. **Robson, B. and Osguthorpe, D. J.,** *J. Mol. Biol.*, 132, 19, 1979.
118. **Nemethy, G. and Scheraga, H. A.,** *Q. Res. Biophys.*, 10, 239, 1977.
119. **Rossky, P. J., Karplus, M., and Rahman, A.,** *Biopolymers*, 18, 825, 1979.
120. **Gibson, K. D. and Scheraga, H. A.,** *Proc. Natl. Acad. Sci.*, 58, 420, 1967.
121. **Hopfinger, A. J.,** *Macromolecules*, 4, 731, 1971.
122. **Pullman, B. P. and Berthod, H.,** *Theor. Chim. Acta*, 36, 317, 1975.
123. **Krimm, S. and Venkatachalam, C. M.,** *Proc. Natl. Acad. Sci.*, 68, 2468, 1971.
124. **Robson, B.,** in *Biophysics of Water*, Franks, F. and Mathias, S., Eds., John Wiley & Sons, 1982.
125. **Renugopalakrishnan, V., Nir, S., and Swissler, T. J.,** *Environmental Effects on Molecular Structure and Properties*, Reidel, Dorbecht, 1976, 109.
126. **Rein, R., Renugopalakrishnan, V., Nir, S., and Swissler, T. J.,** *Int. J. Quantum Chem. Biol. Symp.*, 2, 99, 1975.
127. **Krohn, A.,** *Biochem. Soc. Trans.*, 10, 309, 1982.