Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins[1]

Seiji Tanaka[2a] and Harold A. Scheraga* [2b]

Department of Chemistry, Cornell University, Ithaca, New York 14853.
Received March 18, 1976

ABSTRACT: In a previous paper, a hypothesis for protein folding was proposed, in which the native structure is formed by a three-step mechanism: (A) formation of ordered backbone structures by short-range interactions, (B) formation of small contact regions by medium-range interactions, and (C) association of the small contact regions into the native structure by long-range interactions. In this paper, the *empirical* interaction parameters, used as a measure of the medium- and long-range interactions (the standard free energy, $\Delta G^\circ_{k,l}$, of formation of a contact between amino acids of species k and l) that include the role of the solvent (water) and determine the conformation of a protein in steps B and C, are evaluated from the frequency of contacts in the x-ray structures of native proteins. The numerical values of $\Delta G^\circ_{k,l}$ for all possible pairs of the 20 naturally occurring amino acids are presented. Contacts between highly nonpolar side chains of amino acids such as Ile, Phe, Trp, and Leu are shown quantitatively to be stable. On the contrary, contacts involving polar side chains of amino acids such as Ser, Asp, Lys, and Glu are significantly less stable. While this implies, in a quantitative manner, that it is generally more favorable for nonpolar groups to lie in the interior of the protein molecule and for the polar side chains to be exposed to the solvent (water) rather than to form contacts with other amino acids, many exceptions to this generalization are observed.

In a previous paper,[3] a hypothesis was proposed for protein folding, wherein the globular structure of the native protein is assumed to form by the following mechanism that involves three steps (which *may* proceed simultaneously): (A) Because of short-range interactions,[4] ordered backbone structures, such as α-helical, extended, and chain-reversal conformations, are formed in a system at equilibrium under given conditions (e.g., above the denaturation temperature). (B) When these physical conditions are changed (for example, by changing the temperature and/or solvent composition), so as to introduce medium-range interactions, the equilibrium is shifted, and small contact regions (defined in section II of ref 3), involving medium-range interactions, are nucleated among amino acid residues both in the ordered and in the unordered structures. In this step, the ordered backbone structures formed in step A may be rearranged to some extent to form stable intermediate structures in these contact regions. (C) Finally, the small contact regions formed in the intermediate structures in step B associate, in response to long-range interactions, with possible further small rearrangements of the intermediate structures formed in steps A and B.

The conformations formed in step A were described in terms of statistical mechanical treatments of one-dimensional short-range interaction models.[7-12] To describe the conformations formed in steps B and C, a quantitative measure of the medium- and long-range interactions [that also includes the role of the solvent (water)] is required. For this purpose, in this paper we obtain such a quantitative measure of the medium- and long-range interactions in terms of an *empirical* standard free energy of formation of contacts between pairs of amino acids. A brief description of the method for evaluating these interaction parameters was reported in our previous paper,[3] and we present a more detailed description here. Using these empirical free energies and a Monte Carlo simulation of protein folding [applied to bovine pancreatic trypsin inhibitor, (BPTI)][3] it was demonstrated that the three-step mechanism is required to obtain the globular structure of the native protein.

The Monte Carlo procedure used in ref 3 differs from the model building approach of Ptitsyn and Rashin,[13] from the molecular dynamic treatment of Levitt and Warshel,[14] and from the conformational energy minimization calculation of Burgess and Scheraga,[15] all three of which were applied to the protein folding problem. The Monte Carlo procedure has been applied to many equilibrium and dynamical problems of polymer chains but, as far as we are aware, not heretofore to the protein folding problem.

In section I of this paper, a discussion of the statistical mechanical basis for obtaining parameters for medium- and long-range interactions is presented. In section II, a method to evaluate an *empirical* standard free energy of formation of contacts between pairs of amino acids will be described. In section III, the numerical results will be presented and discussed.

I. Statistical Thermodynamic Discussion of the Parameters for Medium- and Long-Range Interactions between Amino Acids in Proteins

An aqueous solution of a native protein is assumed to be a system involving an equilibrium between an unfolded form, in which all parts of the chain are exposed to the solvent, and a folded form, in which many groups are in contact in the interior and are partially or completely shielded from the solvent. Species of intermediate degrees of folding are also assumed to be involved in this equilibrium. In the Monte Carlo simulation of protein folding,[3] the folding process proceeded because the equilibrium was shifted from the unfolded to the folded form. The intermediates in the Monte Carlo simulation were also assumed to exist in equilibrium with molecules in various states of folding, as in reversible thermal denaturation, and the equilibrium is shifted toward the folded form by lowering the temperature or by changing the solvent composition in stages. The intermediates form as medium- and long-range interactions come into play.

As a first approximation, the medium- and long-range interactions that play a role in steps B and C are taken as those between atoms or groups of atoms (e.g., between side chains or between whole residues) that are spatially close to each other (the criterion for closeness will be specified in section II). We use the term "contact" to designate such spatially close interactions. We neglect the medium- and long-range interactions between atoms or groups of atoms that are *not* spatially close to each other (primarily because the high dielectric constant of water reduces the electrostatic interactions between charged atoms or groups, which are the main contributions to medium- and long-range interactions, and the nonbonded van der Waals interactions are of short-range order). The model, to which this approximation is applied, will

be designated as a "nearest-neighbor three-dimensional interaction model",[16] in the sense that only three-dimensional nearest-neighbor (or first layer) interactions between amino acids are taken into account (within the range of the first hydration shell around each residue, as will be specified in section II), and interactions between residues further away are neglected.

We introduce a further approximation in the nearest-neighbor three-dimensional interaction model by treating spatially close interactions between more than two amino acids as a sum of pairwise ones. For example, if three amino acids (A_1, A_2, and A_3) approach each other closely, we assume that the difference in standard free energy, $\Delta G^\circ_{A_1,A_2,A_3}$, between the separated amino acids and the "ternary complex" may be written as

$$\Delta G^\circ_{A_1,A_2,A_3} = \Delta G^\circ_{A_1,A_2} + \Delta G^\circ_{A_2,A_3} + \Delta G^\circ_{A_3,A_1} \quad (1)$$

In this approximation, we neglect three-body terms and the contribution from "extra" solvent (water) molecules, where the term "extra" refers to those water molecules not counted in those excluded by a contact between each pair of amino acids (A_1, A_2; A_2, A_3; and A_3, A_1). In general, we assume that the standard free energy of formation of any complex of more than two amino acids can be approximated by a sum of pairwise contact standard free energies, i.e.,

$$\Delta G^\circ_{A_1,A_2,A_3,\ldots} = \Delta G^\circ_{A_1,A_2} + \Delta G^\circ_{A_2,A_3} + \ldots \quad (2)$$

Thus, the association of amino acids (or the formation of a contact between amino acids) in steps B and C can be represented as

$$\begin{aligned} A_1 + A_1 &\rightleftharpoons A_{1,1} \\ A_1 + A_2 &\rightleftharpoons A_{1,2} \\ &\vdots \\ A_k + A_l &\rightleftharpoons A_{k,l} \\ &\vdots \end{aligned} \quad (3)$$

where $A_{k,l}$ includes not only binary complexes (a contact between two amino acids) between residues A_k and A_l but also higher complexes which can be thought of as made up of pairwise ones between A_k and A_l.

On the basis of the above assumptions, all association reactions are reduced to pairwise ones, i.e., as a general form of eq 3,

$$A_k + A_l \rightleftharpoons A_{k,l} \quad (4)$$

where A_k and A_l designate species that are not in contact, and $A_{k,l}$ designates species that are. Thus, since, at any stage of folding, a protein is considered as a system in equilibrium, involving many *simultaneous* equilibria of the type of eq 4 (between all possible pairs of the 20 naturally occurring amino acid residues), we may assign an equilibrium constant $K_{k,l}$ defined as

$$K_{k,l} = X_{k,l}/X_k X_l \quad (5)$$

where the X's are the mole fractions of the species in eq 4. The corresponding standard free energy change is

$$\Delta G^\circ_{k,l} = -RT \ln K_{k,l} \quad (6)$$

This standard free energy change, in principle, can be computed from the partition functions of the species involved, viz., $Q_{k,l}$, Q_k, and Q_l. These partition functions, in turn, are calculable theoretically from appropriate potential functions (including the effect of the solvent by some appropriate approximation, such as the solvent-shell model[17]). However, the parameters involved in such a solvent-shell model are very approximate, and efforts are currently under way in our laboratory to improve them, using a Monte Carlo treatment of the solvation of small-molecule solutes in water.

Alternatively, the standard free energy of association of amino acids may be obtained experimentally by studying association reactions of amino acids in aqueous solution, as represented by eq 4 and 5. However, this would require a very extensive set of experiments on *pair* association (not the multiple association involved in solubility studies) between the 20 naturally occurring amino acids.

Pending the acquistion of such hydration free energies, we have resorted[3] to a temporary alternative approach to obtain empirical values of $\Delta G^\circ_{k,l}$ from another type of experimental data on $K_{k,l}$, in order to make progress in circumventing the most difficult obstacle in protein folding, viz., the existence of many minima on the multidimensional conformational energy surface. Thus, for the present, we have resorted[3] to a crude but simple model to treat medium- and long-range interactions to solve the multiple-minimum problem. The medium- and long-range interaction model can be improved subsequently, when the Monte Carlo calculations on aqueous solutions are completed; i.e., if the multiple-minimum problem can be solved by using a simple model to introduce medium- and long-range interactions, the parameters for the latter can always be improved in a subsequent stage. In section II, we present a method for evaluating empirical parameters to describe the medium- and long-range interactions.

II. Empirical Methods

Contacts between amino acid residues of a protein that approach each other arise because of medium- and long-range interactions in steps B and C.[3] The solvent (water) can play an important role in stabilizing the conformation of the protein (and, hence, in forming such contacts). We, therefore, describe such a specific local interaction as a *contact* between two amino acids (or between the side chains and/or backbone groups of two amino acids). Such a specific local interaction is said to exist between the two groups, A and B, in water if these groups are in contact, or, more quantitatively, if the distance between them, r_{AB}, satisfies the relation

$$r_A^{(w)} + r_B^{(w)} \leq r_{AB} < r_A^{(w)} + r_B^{(w)} + 2r_{H_2O}^{(w)} \quad (7)$$

where $r_A^{(w)}$, $r_B^{(w)}$, and $r_{H_2O}^{(w)}$ are the van der Waals radii of groups A, B, and a water molecule, respectively [$r_{H_2O}^{(w)}$ is taken as 1.40 Å]. The moieties A and B may be considered to be individual atoms or groups of atoms (side-chain and backbone groups) or amino acid residues. For reasons cited in section I, the local interaction between A and B is neglected when

$$r_{AB} \geq r_A^{(w)} + r_B^{(w)} + 2r_{H_2O}^{(w)} \quad (8)$$

Infinite repulsion arises (excluded volume effect) when

$$r_{AB} < r_A^{(w)} + r_B^{(w)} \quad (9)$$

Thus, the three-dimensional structure of a protein can be represented symbolically by the presence or absence of contacts between the ith and jth residues [$1 \leq (i, j) \leq N$ and $i < j$, where N is the chain length], when account is taken of the chain connectivity. For example, contact maps of the three-dimensional structures of several proteins are shown in Figure 1 of ref 3 (BPTI) and in Figures 2–4 of ref 5 (rubredoxin, ferricytochrome c, and lysozyme). In these contact maps, a contact is said to exist if at least one pair of atoms (one atom in the ith and one in the jth residues) satisfies eq 7. The van der Waals distances[18,19] for a pair of atoms or groups of atoms used in this paper are listed in Table I. In this analysis, the contacts considered are between the side chains of the 19 non-glycine residues (or the full residue, in the case of glycine) in the ith position and the side chains of the 19 non-glycine

Table I
van der Waals Contact Distances for Pairs of Atoms and Groups of Atoms[a]

	van der Waals contact distance, Å							
	C′	O	N[b]	CH[c]	C_{ar}	NH_2	OH	S
C′	2.9	2.7	2.7	3.2	3.1	3.1	3.0	3.30
O		2.6	2.6	3.0	3.0	3.0	2.9	3.15
N[b]			2.6	3.0	3.0	3.0	2.9	3.15
CH[c]				3.5	3.4	3.4	3.3	3.60
C_{ar}					3.4	3.4	3.3	3.55
NH_2						3.4	3.3	3.55
OH							3.2	3.45
S								3.70

[a] The values for S are cited from ref 18 and all other values from ref 19. [b] These values were used for N and backbone NH. [c] These values were used for CH, CH_2, and CH_3.

residues (or the full residue, in the case of glycine) in the jth position. Also, since our present purpose is to obtain parameters that describe the medium- and long-range interactions in a protein, we do not consider short-range contacts;[4,20] i.e., in the present analysis, contacts from residue i to residues $i + 1$, $i + 2$, $i + 3$, and $i + 4$ are omitted.

The same analysis, as mentioned above, was made for 25 proteins,[21] whose x-ray structures are known; they are myoglobin, lysozyme, ribonuclease S, deoxyhaemoglobin α chain, deoxyhaemoglobin β chain (α and β chains are counted as two proteins), α-chymotrypsin (B and C chains are counted as one protein), carboxypeptidase A, subtilisin BPN′, tosyl elastase, staphylococcal nuclease, papain, ferricytochrome c, lactate dehydrogenase, cytochrome b_5, thermolysin, concanavalin, carp myogen, sea lamprey haemoglobin, rubredoxin, ferredoxin, trypsin, pancreatic trypsin inhibitor, glyceraldehyde phosphate dehydrogenase, clostridial flavodoxin, and high potential iron protein.

With the above definition of a contact, the data for these 25 proteins were used to obtain the number, $N_{(n)k}$, of the amino acid of type A_k with no contacts, and the number, $N_{k,l}$, of complexes, $A_{k,l}$, with two residues, A_k and A_l, in contact. The subscript (n) emphasizes that the amino acid of type A_k has "no" contact. The contacts considered here are side chain–side chain, side chain–glycine, and glycine–glycine contacts. Since any given amino acid residue, A_k, in a protein can form a contact with more than one other residue, the system may be represented in terms of the 210 *simultaneous* equilibria, involving the 20 species of the naturally occurring amino acids, shown in eq 3. Each of the equilibria in eq 3 has an equilibrium constant given by eq 5, where we may define the following quantities that resemble the mole fractions of eq 5:

$$X_k = N_{(n)k}/N_T$$

$$X_l = N_{(n)l}/N_T$$

and

$$X_{k,l} = N_{k,l}/N_T \tag{10}$$

where the total number of complexes and amino acids without contacts, N_T, in the system is given by

$$N_T = \sum_{k=1}^{20} N_{(n)k} + \sum_{k \geq l=1}^{20} N_{k,l} \tag{11}$$

From the *empirical* equilibrium constants of eq 5 (using X_k, X_l, and $X_{k,l}$ defined in eq 10), we obtain the *empirical* standard free energy of formation of a contact (or of a complex $A_{k,l}$) between amino acids A_k and A_l by means of eq 6. In obtaining *empirical* values of $K_{k,l}$ from the data of Table II, we use the "mole fractions" directly, with no introduction of activity coefficients.

III. Results and Discussion

Using the x-ray coordinates for 25 proteins,[21] an analysis was made of the number of times, $N_{(n)k}$, that an amino acid of type A_k had no contact and of the number of times, $N_{k,l}$, that amino acid A_k formed a contact with amino acid A_l (or the number of complexes, $A_{k,l}$); these data are summarized in Table II. For glycine, we considered the whole residue; hence, the numbers $N_{k,l}$ in the row and column for Gly in Table II indicate the number of contacts between a whole glycine residue and the side chains of all other 19 non-glycine residues and between two whole glycine residues. For the reason cited in section II, short-range contacts between residue i and residues $i + 1$, $i + 2$, $i + 3$, and $i + 4$ were not included in $N_{k,l}$. Thus, to obtain the values of $N_{(n)k}$ and $N_{k,l}$ in Table II, we used the van der Waals distances of Table I and the criteria of eq 7–9 to determine whether two atoms were in contact (one atom in the kth side chain and one in the lth side chain; or the kth and lth full residues in the case of glycine).

From the values of $N_{(n)k}$ and $N_{k,l}$ in Table II, the *empirical* standard free energies, $\Delta G°_{k,l}$, were computed (for $k \geq l = 1$ to 20). The results for $\Delta G°_{k,l}$, calculated at 30 °C, are given in Table III.[22–24]

According to the foregoing formulation, the values of $\Delta G°_{k,l}$ do *not* depend on the total number of occurrences of each of the amino acids. Thus, these values do *not* reflect the fact that some amino acids occur more frequently in proteins than do others (except for the fact that the data are statistically more reliable for the more frequently occurring residues). While eq 5 and 10 might seem to imply that $K_{k,l}$ is proportional to N_T, actually N_T is itself a function of the $N_{(n)k}$'s and $N_{k,l}$'s (see eq 11). Therefore, as the protein data set increases in size (with a resulting increase in N_T) it does not necessarily follow that $K_{k,l}$ increases correspondingly. In all likelihood, the values of $K_{k,l}$ converge for a sufficiently large data set, and it is reasonable to expect that the 25 proteins considered here constitute a sufficiently large set for such convergence to have occurred.

As seen in Table III, the most stable contact is that between Ile and Ile side chains ($\Delta G°_{k,l} = -8.2$ kcal/mol). Other stable contacts involve the nonpolar side chains, Ile···Phe, Ile···Trp, and Ile···Leu ($\Delta G°_{k,l} = -8.0$, -7.8, and -7.5 kcal/mol, respectively). These favorable contacts undoubtedly are stabilized by hydrophobic interactions. In fact, as can be seen in Table III, the contacts between nonpolar side chains other than those mentioned above are relatively stable, e.g., Ile···Met, Phe···Trp, Ile···Tyr, Ile···Val, Cys···Ile, and Phe···Phe ($\Delta G°_{k,l} = -7.4$, -7.4, -7.4, -7.3, -7.3, and 7.1 kcal/mol, respectively).

On the other hand, the amino acids with polar side chains have a stronger preference to be exposed to the solvent (water); in other words, the contacts involving polar amino acids (especially between polar side chains with the same sign of charge) are not as stable as those involving nonpolar side chains, e.g., Ser···Ser, Asp···Asp, Lys···Lys, Asp···Ser, Glu···Glu and Lys···Ser ($\Delta G°_{k,l} = -2.5$, -2.7, -2.7, -2.7, -2.8, and -3.0 kcal/mole, respectively). However, as pointed out from an examination of contact maps,[5] most contact regions are *not* stabilized only by contacts between nonpolar side chains. Thus, it is of interest to note that (i) a contact between polar side chains with the same sign of charge can form, even though it is not as stable as those between nonpolar side chains (e.g., Arg···Lys, His···Lys, Arg···Arg, Arg···His, and His···His, with $\Delta G°_{k,l} = -3.6$, -4.1, -4.3, -4.9, and -4.9 kcal/mol, respectively), and (ii) a contact between polar side chains of opposite charges (e.g., Asp···Lys, Glu···Lys, Arg···Asp, Arg···Glu, Asp···His, and Glu···His, with $\Delta G°_{k,l} = -3.2$, -3.8, -3.9, -4.2, -4.3, and -4.5 kcal/mol, respectively) is not as stable as those between nonpolar side chains. Presumably, screening effects

Table II: Number of Contacts between Side Chains of Amino Acids and Number of Noncontacts of Amino Acid Side Chains

	Number of contacts $N_{k,l}$																			No. of non-contacts $N_{(n)k}$
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly[a]	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	
Ala	36	20	39	36	15	34	27	(115)	25	119	174	55	26	62	33	64	70	38	65	101	70
Arg		14	34	35	13	29	34	(58)	17	33	47	20	9	32	20	44	36	15	45	44	11
Asn			22	31	21	30	27	(80)	25	51	63	39	7	27	22	42	46	21	50	50	34
Asp				24	10	17	38	(83)	32	43	60	54	13	32	23	37	38	24	52	53	54
Cys					48	13	9	(53)	11	33	49	14	11	32	24	28	22	17	43	53	2
Gln						7	19	(75)	17	29	53	32	1	21	23	44	28	22	32	46	16
Glu							9	(54)	24	35	53	70	12	22	17	41	28	15	40	39	30
Gly[a]								(75)	(40)	(109)	(155)	(92)	(32)	(67)	(62)	(145)	(116)	(52)	(103)	(183)	34
His									8	27	49	21	7	38	15	43	36	23	44	42	5
Ile										81	230	53	44	103	41	65	67	40	81	216	1
Leu											167	86	58	175	46	93	110	68	105	306	9
Lys												20	21	32	34	51	50	20	64	67	49
Met													6	20	16	14	16	17	19	52	2
Phe														53	19	36	39	41	45	114	2
Pro															11	38	32	19	42	58	19
Ser																34	69	31	71	106	77
Thr																	29	20	53	79	42
Trp																		7	32	62	1
Tyr																			33	85	4
Val																				152	13

[a] For glycine, the full residue is taken and its contact with 19 other side chains and glycine counted. The values for glycine are enclosed in parentheses.

Table III: Standard Free Energy of Formation of Contact between Side Chains of Amino Acids[a] $\Delta G^{\circ}_{k,l}$

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly[b]	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	−2.6	−3.4	−3.1	−2.8	−4.2	−3.5	−3.0	(−3.8)	−4.0	−5.9	−4.8	−3.1	−4.6	−5.1	−3.4	−2.9	−3.3	−5.2	−4.7	−4.3
Arg		−4.3	−4.1	−3.9	−5.3	−4.5	−4.2	(−4.5)	−4.9	−6.2	−5.1	−3.6	−5.0	−5.8	−4.2	−3.8	−4.0	−5.8	−5.6	−4.9
Asn			−3.2	−3.1	−4.9	−3.8	−3.4	(−4.0)	−4.4	−5.8	−4.6	−3.3	−4.2	−5.0	−3.6	−3.1	−3.5	−5.3	−5.0	−4.3
Asp				−2.7	−4.2	−3.2	−3.3	(−3.7)	−4.3	−5.4	−4.3	−3.2	−4.3	−4.9	−3.3	−2.7	−3.1	−5.1	−4.7	−4.0
Cys					−7.1	−5.0	−4.4	(−5.4)	−5.6	−7.3	−6.2	−4.4	−6.2	−6.8	−5.3	−4.6	−4.8	−6.9	−6.6	−6.0
Gln						−3.4	−3.6	(−4.4)	−4.7	−5.9	−5.0	−3.7	−3.5	−5.3	−4.0	−3.6	−3.7	−5.8	−5.2	−4.7
Glu							−2.8	(−3.8)	−4.5	−5.7	−4.6	−3.8	−4.6	−5.0	−3.5	−3.2	−3.3	−5.2	−4.9	−4.2
Gly[b]								(−3.9)	(−4.7)	(−6.3)	(−5.2)	(−3.8)	(−5.1)	(−5.6)	(−4.2)	(−3.8)	(−4.1)	(−5.8)	(−5.4)	(−5.1)
His									−4.9	−6.6	−5.6	−4.1	−5.4	−6.4	−4.5	−4.3	−4.5	−6.5	−6.1	−5.3
Ile										−8.2	−7.5	−5.6	−7.4	−8.0	−6.0	−5.5	−5.9	−7.8	−7.4	−7.3
Leu											−6.0	−4.6	−6.3	−7.0	−4.8	−4.4	−4.8	−6.8	−6.2	−6.2
Lys												−2.7	−4.7	−4.9	−3.6	−3.0	−3.3	−5.0	−4.9	−4.2
Met													−5.8	−6.6	−5.1	−4.1	−4.6	−6.9	−6.1	−6.0
Phe														−7.1	−5.2	−4.7	−5.1	−7.4	−6.6	−6.5
Pro															−3.5	−3.4	−3.6	−5.6	−5.2	−4.7
Ser																−2.5	−3.3	−5.0	−4.7	−4.2
Thr																	−3.1	−5.1	−4.9	−4.4
Trp																		−6.8	−6.8	−6.5
Tyr																			−6.0	−5.9
Val																				−5.5

[a] In units of (kcal/mol) at 30 °C. [b] For glycine, the full residue is taken, and its contact free energies with 19 other side chains and glycine are evaluated; glycine also serves as a prototype of the backbone for non-glycine residues (as in ref 3).

reduce the repulsive forces in type (i) contacts and weaken the attractive forces in type (ii) contacts, with an interplay between electrostatic and other forces. In this connection, it should be kept in mind that it is possible to form a contact between the nonpolar parts of polar side chains.[25] Thus, the chain connectivity of a protein may bring about type (i) and (ii) contacts, when neighboring nonpolar side chains can form stable contacts with the nonpolar portions of polar residues to overcome the repulsive forces between side chains of the same sign of charge.

It should be emphasized that the data of Table III imply more than the validity of the generally held qualitative view that nonpolar residues tend to occur in the interior of a protein, while polar ones are on the exterior. As pointed out in ref 5, nonpolar residues can exist on the exterior and polar ones in the interior; this is reflected in the data of Table III.

As discussed in section I, the standard free energy of formation of a contact between amino acid residues in proteins is obtainable, in principle, from theoretical calculations or from experimental measurements. In section II, we presented an alternative approach for evaluating a set of parameters to represent medium- and long-range interactions by using as simple a model as possible. The simplicity of the model (e.g., a pairwise representation of the interactions between amino acids) enables the use of the interaction parameters to simulate protein folding (see next paragraph) by substantially reducing the required computer time to an extent that makes the computation practicable. Even within the context of the present model (described in section II), it would be possible to introduce more physically meaningful improvements, such as the following two features. (i) As mentioned in section I, the interaction parameters can be represented in terms of interactions between small units (say, atoms or groups of atoms smaller than the side chains and backbone employed in this paper). Such an improvement would provide a more precise treatment of a contact point; i.e., the magnitude of the interaction would then depend on the number and types of atoms that are in contact. However, such an improvement would be achieved at the cost of increased computer time. (ii) As an approximation, we used the interaction parameter between two glycine residues as a prototype of that between two backbone groups of non-glycine residues, relying on the fact that the atomic species that make up a glycine residue differ only by a hydrogen atom on the α carbon from the backbone of a non-glycine residue. This is an approximation to the extent that the side chain of a non-glycine residue would influence the location of its backbone within the protein molecule. It is possible to improve the treatment of backbone–backbone and side chain–backbone interactions by analyzing x-ray data in a manner similar to that described for side chain–side chain, side chain–glycine, and glycine–glycine interactions in section II.

The *empirical* standard free energy of formation of a contact between amino acids of species A_k and A_l, $\Delta G°_{k,l}$, given in Table III, can be used as an interaction parameter to account for the medium- and long-range interactions [including the role of the solvent (water)] that determine the conformations of a protein in steps B and C of the three-step mechanism[3] of protein folding. In computations on protein folding, $\Delta G°_{k,l}$ for the glycine residue (where A_k and/or A_l is Gly) can also serve as a prototype for the backbone of non-glycine residues, when considering contacts involving these backbones.[22] The values of $\Delta G°_{k,l}$ have been used[3] in this manner, in a simulation of protein folding by a Monte Carlo procedure, in which the ordered conformation of step A was taken from x-ray information and steps B and C were simulated.[24] (The three-step mechanism was also used[5] to detect possible pathways of protein folding.)

The data of Table II were taken from x-ray crystal data obtained under a variety of conditions. In using them, we assume that all such crystal structures correspond to the structures of native proteins in water at some undefined temperature, which we have arbitrarily taken as 30 °C. Thus, when the free energy of formation of the native protein is computed from the data of Table III, it pertains to this temperature. However, in using the data of Table III in the Monte Carlo calculation,[3] in which the temperature varies in stages from the (high) denaturation temperature to 30 °C, as the protein passes through successive (assumed) equilibrium states, we have neglected the temperature dependence of the parameters, hence of $\Delta G°_{k,l}$, simply because the computation of the temperature dependence of the parameters is a formidable (but by no means insoluble) task. As far as we can determine, Ptitsyn and Rashin[13] and Levitt and Warshel[14] also assumed that their parameters are independent of temperature (no doubt for the same reason).

A complete simulation of protein folding is in progress,[26] whereby statistical mechanical treatments of protein conformation (based on short-range interaction models)[4,7-12,27] are incorporated into a Monte Carlo procedure,[3] as had been done previously[28] in treating the three-dimensional structure of a polymer molecule.

References and Notes

(1) This work was supported by research grants from the National Institute of General Medical Sciences, National Institutes of Health, U.S. Public Health Service (GM-14312), and from the National Science Foundation (BMS75-08691).

(2) (a) From Kyoto University, 1972–1975; (b) to whom requests for reprints should be addressed.

(3) S. Tanaka and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 3802 (1975).

(4) In this paper, as well as in ref 3 and 5, the terms short-, medium-, and long-range interactions are defined as follows: if i and j refer to the positions of two residues in the amino acid sequence, then these terms pertain to interactions for which $|i - j| \leq 4$, 5 to 20, and >20, respectively. This differs from the nomenclature of ref 6, where medium range pertained to $1 \leq |i - j| \leq 4$.

(5) S. Tanaka and H. A. Scheraga, *Macromolecules*, submitted (paper on a hypothesis about the mechanism of protein folding).

(6) P. K. Ponnuswamy, P. K. Warme, and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, **70**, 830 (1973).

(7) S. Tanaka and H. A. Scheraga, *Macromolecules*, **9**, 142 (1976): paper 1.

(8) S. Tanaka and H. A. Scheraga, *Macromolecules*, **9**, 159 (1976): paper 2.

(9) S. Tanaka and H. A. Scheraga, *Macromolecules*, **9**, 168 (1976): paper 3.

(10) S. Tanaka and H. A. Scheraga, *Macromolecules*, **9**, 812 (1976): paper 4.

(11) S. Tanaka and H. A. Scheraga, *Macromolecules*, in press: paper 5.

(12) S. Tanaka and H. A. Scheraga, *Macromolecules*, submitted: paper 6.

(13) O. B. Ptitsyn and A. A. Rashin, *Biophys. Chem.*, **3**, 1 (1975).

(14) M. Levitt and A. Warshel, *Nature (London)*, **253**, 694 (1975).

(15) A. W. Burgess and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 1221 (1975).

(16) We use the term "nearest-neighbor three-dimensional interaction model" to distinguish the interactions described in the text from those in the short-range interaction model[4] where the range of interaction pertains to the one dimensionality (along the chain) that was treated in our previous papers.[7-12]

(17) K. D. Gibson and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, **58**, 420 (1967).

(18) L. Pauling, "The Nature of the Chemical Bond," 3d ed, Cornell University Press, Ithaca, N.Y., 1960, p 260.

(19) S. J. Leach, G. Nemethy, and H. A. Scheraga, *Biopolymers*, **4**, 369 (1966).

(20) The attractive forces that lead to the formation of a contact between amino acids should be regarded as the sum of all contributions from nonbonded and electrostatic interactions and from solvation (i.e., the free energy for removing water molecules from the first hydration shell around an atom[17]) for all atoms which are in contact. The criterion for a contact, given by eq 7, allows the two groups to be separated by up to 2.8 Å (the diameter of a water molecule) and still be subject to attractive forces.

(21) References to the original papers, in which the x-ray data for these 25 proteins are reported, are given in the footnotes of Table I of ref 10.

(22) The data in Table III may also be used for interactions involving the backbones of non-glycine residues. For this purpose, the whole glycine residue is taken as a prototype of the backbone of the non-glycine residue. For example, $\Delta G° = -4.5$ kcal/mol for the interaction of an Arg side chain with the backbone of *any* other residue.

(23) An earlier form[24] of the data of Table III was used to calculate free energies in the Monte Carlo procedure described in ref 3. In this Monte Carlo simulation, in which the backbone and side chains were represented by

spheres, the spherical representation was used *only* to test for contacts; when evaluating the free energy of contacts (as in Figure 3 of ref 3), the spherical representation was abandoned, and the earlier form[24] of the data of Table III was used.[3]

(24) In our earlier calculations,[3] we used numerical values of $\Delta G°_{k,l}$ that were evaluated on the basis of a different definition from that described in section II of this paper (but using the same values of $N_{(n)k}$ and $N_{k,l}$ given in Table II). Therefore, the effect of this altered definition of $\Delta G°_{k,l}$ on the previous results[3] was checked. It was found that the general appearances of the contact regions in steps B and C are the same, but the precise locations of the contacts within each region and the type of contact (represented by numerals in the squares on the contact maps of Figures 4 and 5 of ref 3) are altered somewhat. Thus, the most important implication, discussed in the last paragraph of ref 3, is unchanged. The main reason for this is that both the old and the new sets of values of $\Delta G°_{k,l}$ reflect the general tendencies for strong contacts between nonpolar residues and for exposure of polar ones to the solvent.

(25) G. Nemethy, I. Z. Steinberg, and H. A. Scheraga, *Biopolymers*, **1**, 43 (1963).
(26) S. Tanaka and H. A. Scheraga, work in progress.
(27) S. Tanaka and H. A. Scheraga, *Macromolecules*, **8**, 494 (1975).
(28) S. Tanaka and A. Nakajima, *Macromolecules*, **5**, 708, 714 (1972).

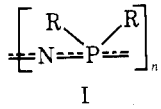
Conformational Analysis of Poly(dihalophosphazenes)[1,2]

H. R. Allcock,* R. W. Allen, and J. J. Meister

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802. *Received September 26, 1975*

ABSTRACT: The poly(dihalophosphazenes), $(NPF_2)_n$, $(NPCl_2)_n$, and $(NPBr_2)_n$, have been examined by conformational analysis using nonbonding intramolecular interactions based on a 6:12 Lennard-Jones potential and a Coulombic term. The results provide an insight into the reasons for the low glass transition temperatures, the high chain flexibilities, and the conformational preferences of these molecules.

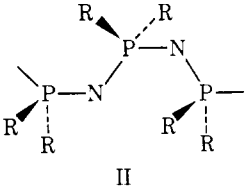
High molecular weight polyphosphazenes, of general formula I, have physical properties that are unusual and, in many



I

cases, unexpected.[3-16] The three halogeno derivatives, $(NPF_2)_n$, $(NPCl_2)_n$, and $(NPBr_2)_n$, are rubbery, elastomeric materials over a broad temperature range. For the fluoro derivative the elastomeric properties are maintained from the glass transition at −95 to +270 °C.[12] The corresponding rubbery range for the chloro derivative is from −63 to +350 °C and for the bromo compound from −15 to +270 °C. The methoxy and ethoxy derivatives,[3,4] $[NP(OCH_3)_2]_n$ and $[NP(OC_2H_5)_2]_n$, are flexible, film-forming materials with glass transition temperatures near −80 °C. These facts suggest that the polyphosphazene backbone has an unusually high torsional mobility.

The conformational properties of polyphosphazenes are also unusual. The polymers, $[NPCl_2]_n$, $[NP(OCH_2CF_3)_2]_n$, $[NP(OCH_2C_3F_5)_2]_n$, $[NP(OC_6H_5)_2]_n$, and $[NP(OC_6H_4Cl\text{-}p)_2]_n$, crystallize when oriented. The x-ray diffraction patterns suggest a chain repeating distance of ~4.9 Å which can be rationalized in terms of a cis-trans planar arrangement (II).[3,4,17-19] On the other hand, poly(difluorophosphazene),



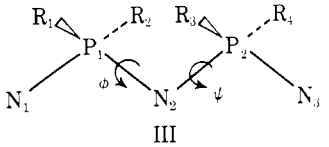
II

$(NPF_2)_n$, shows two different x-ray patterns, one at temperatures between 25 and −36 °C, indicative of a 6.45 Å chain repeat, and the other at temperatures below −56 °C, with a 4.86 Å repeat distance and a cis-trans planar conformation.[12] The bromo derivative crystallizes only with difficulty,[19] but similarities have been reported between the x-ray patterns of $(NPBr_2)_n$ and $(NPCl_2)_n$.[20] It was of some interest to deduce the reasons for these conformational characteristics.

In this paper we have attempted to explain these experimental facts with the use of nonbonding intramolecular potential energy calculations. This paper deals with the three simplest systems, $(NPF_2)_n$, $(NPCl_2)_n$, $(NPBr_2)_n$, and the yet unsynthesized $(NPI_2)_n$. The accompanying paper considers organo-phosphazene polymers.

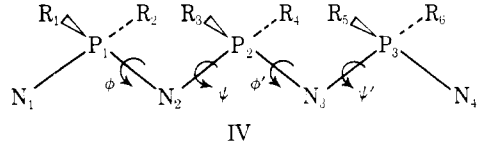
Model and Method

The Model. Two structural alternatives were employed which made use of the short chain segments shown in III and



III

IV for calculation of the nonbonding interactions for independent incremental torsion of bonds P_1–N_2, N_2–P_2, P_2–N_3, and N_3–P_3. The initial position for the computations was the trans-trans conformation, shown in III and IV.



IV

The choice of bond angles and bond lengths was based on an analysis of published x-ray crystallographic data for cyclic trimers, tetramers, a pentamer,[21-28] and, in the case of $(NPF_2)_n$, the high polymer.[12] It was recognized that the restrictions imposed by six- or eight-membered ring structures could modify the preferred skeletal angles or interatomic distances. Therefore, trial models were used that explored changes in bond angles or bond lengths for values above and below those found for the small molecules. Table I lists the structural parameters used. The comprehensive background of crystallographic data for cyclic phosphazenes indicates that all the P–N bonds should be of equal length when only one type of R group is present.[21]

For structure III, the coordinate calculation programs computed intramolecular distances between the atoms for 10° incremental torsion of bonds P_1–N_2 and N_2–P_2 through 360°. Structure III was considered a nine-body problem, with neglect of those interactions, such as N_1–R_1, R_2–N_2, P_1–P_2, etc.,