1,4–trans-1,4 (tt), cis-1,4–cis-1,4 (cc), trans-1,4–cis 1,4 (tc), cis-1,4–trans-1,4 (ct), and trans-1,4 –3,4 (t-3,4) were obtained from the ^{13}C spectra based on the assignment in Table III. The results are summarized in Table IV. The fractions of these dyad sequences were calculated assuming Bernoullian statistics. The observed and calculated dyad fractions are in good agreement, indicating that trans-1,4, cis-1,4, and 3,4 units are distributed randomly in poly1PB. In the case of sample no. 14 containing higher 3,4 units, however, the observed fractions of tc, ct, and cc sequences were larger than those calculated. This is due to the contribution of c-3,4-, 3,4-c, and 3,4-t sequences which were not negligible when the 3,4 content increased. The absorptions of carbons in such sequences may have appeared in the region overlapped with peaks E, I, and J.

As to head-to-head or tail-to-tail linkages, the methine carbon resonance in a head-to-head linkage is predicted to have a peak at δ 55.5 ppm according to calculations using the parameters proposed by Conti.[22] There was no observable signal in that region in the ^{13}C NMR spectra of poly1PB. Therefore, poly1PB prepared by anionic polymerization could be considered to have very few head-to-head and consequently tail-to-tail linkages. This shows that the arrangement of head-to-tail linkages is controlled by the living end, and the microstructure of the terminal monomer unit is little affected by the structure of the penultimate unit.

Acknowledgment. This work was partially supported by Grant in aid Ministry of Education, Japan (No. 243019).

References and Notes

(1) K.-F. Elgert, B. Stützel. P. Frezel, and H.-J. Cantow, *Makromol. Chem.*, **170**, 257 (1973); E. R. Santee, Jr., L. O. Malotky, and M. Morton, *Rubber Chem. Technol.*, **46**, 1156 (1973).
(2) Y. Tanaka, H. Sato, M. Ogawa, K. Hatada, and Y. Terawaki, *J. Polym. Sci., Polym. Lett. Ed.*, **12**, 369 (1974).
(3) F. Conti, A. Segre, P. Pini, and L. Porri, *Polymer*, **15**, 5 (1974).
(4) Y. Tanaka, H. Sato, K. Hatada, Y. Terawaki, and H. Okuda, *Makromol. Chem.*, **178**, 1823 (1977).
(5) H. Sato, A. Ono, and Y. Tanaka, *Polymer*, **18**, 580 (1977).
(6) T. Masuda, T. Mori, and T. Higashimura, *J. Polym. Sci., Polym. Chem. Ed.*, **12**, 1385 (1974).
(7) R. Asami, K. Hasegawa, and T. Onoe, *Polym. J.*, **8**, 43 (1976).
(8) R. Asami and H. Murai, *Polym. Prepr. (Jpn.)*, 176 (1967).
(9) E. I. Becker, "Organic Syntheses", Collect. Vol. IV, Wiley, New York, N.Y., 1963, p 771.
(10) W. N. Smith, Jr., *J. Organomet. Chem.*, **82**, 1 (1974).
(11) R. Waack and M. A. Doran, *J. Am. Chem. Soc.*, **91**, 2456 (1969).
(12) F. A. Bovey, G. V. D. Tiers, and G. Filipovich, *J. Polym. Sci.*, **38**, 73 (1959).
(13) M. Morton, R. D. Sanderson, and R. Sakata, *Macromolecules*, **6**, 181 (1973).
(14) M. Morton, R. D. Sanderson, R. Sakata, and L. A. Falvo, *Macromolecules*, **6**, 186 (1973).
(15) A. D. Jenkins, Ed., "Progress in Plymer Science", Vol. 1, Pergamon Press, Elmsford, N.Y., 1967.
(16) H. E. Adams, R. L. Bebb, L. E. Forman, and L. B. Wakefield, *Rubber Chem. Technol.*, **45**, 1252 (1972).
(17) R. Asami, *Kogyo Kagaku Zasshi*, **73**, 2272 (1970).
(18) M. W. Duch and D. M. Grant, *Macromolecules* **3**, 165 (1970).
(19) J. I. Gunneen, G. M. C. Higgins, and W. F. Watson, *J. Polym. Sci.*, **40**, 1 (1959).
(20) Y. Tanaka and H. Sato, *Polymer*, **17**, 113 (1976).
(21) A. L. Segre, M. Delfini, F. Conti, and A. Boicelli, *Polymer*, **16**, 338 (1975).
(22) F. Conti, M. Delfini, and A. L. Segre, *Polymer*, **18**, 310 (1977).
(23) V. D. Mochel, *J. Polym. Sci., Part A-1*, **10**, 1009 (1972).

Chain Dimension and Effective Potential Energy of Globular Proteins

Ken Nishikawa and Tatsuo Ooi*

Institute for Chemical Research, Kyoto University, Uji, Kyoto-Fu 611, Japan.
Received January 5, 1978

ABSTRACT: In this study, a statistical analysis on chain dimensions of irregular parts in native proteins, three-dimensional structures of which are known, has revealed that *short* segments (up to $n = 8$) of the irregular backbone conformations in a protein can be treated, in the sence of a statistical ensemble, as the "unperturbed" (devoid of explicit long-range interactions) polypeptide chain. The single-residue energy primarily responsible for this unperturbed chain dimension is, however, the Pohl's empirical potential energy rather than the atom pair potential energy usually employed. The distinct quantitative difference between the Pohl's energy map and the atom pair energy map is clearly shown in the difference of the characteristic ratios calculated from both energy maps. These evidences indicate that the potential energy for a single residue within a globular protein is modified due to long-range interresidue interactions. Theoretical calculations were carried out on the unperturbed polypeptide chain to obtain the mean squares of end-to-end distances by the method of Flory and distributions of end-to-end distances by the Monte-Carlo technique and served for comparison with the corresponding observed data taken from 19 globular proteins.

Accumulation of high-resolution x-ray data of globular proteins has enabled one to analyze three-dimensional structures of proteins and roles of the constituent amino acid residues in the folding of polypeptide chains by various statistical methods.[1-4] Recently Crippen[5] has presented the dependence of the chain dimension (the mean value of end-to-end distances) of protein segments on the separation of amino acid residues. Also, Monte-Carlo simulations of chain statistics have been carried out for several simple polypeptides.[6-9] The combination of these two methods, i.e., the comparison of the chain dimension observed in globular proteins with that of randomly coiled polypeptides, would provide information about the character of interacting forces within a protein molecule. In our knowledge, there have been few publications of the study along this line; e.g., for a rather limited case of β-turn conformations a comparison of calculated and observed distributions of end-to-end distances has been made.[10]

Although a native protein is by no means a random coil, a large amount of structural data of native proteins may be regarded as a statistical ensemble. The compact shape of a native protein, however, suggests that long-range net attractive forces are dominant in determination of the chain statistics of protein segments. The terms used here, short- and long-range interactions, are atomic interactions within a single residue and all the other interactions between residues, re-

0024-9297/78/2211-0644$01.00/0

spectively, as defined in general polymer theories. A part of the long-range interactions may modify the short-range interaction as additional interactions and the resulting single-residue potential energy, on the average, may be regarded as the effective potential energy of a residue. Of course, it is impossible to include all the long-range interactions in the effective potential since the complicated folding of the chain of a protein may be achieved by long-range interactions. Therefore, we may regard any amino acid residue in a protein obeyed with this effective potential energy together with the inherent long-range interactions not reduced to the single-residue energy.

Since chain statistics of relatively short segments are known to be little affected by the long-range interaction,[11] we pay attention to relatively short segments (say up to 10–20 residues) of native proteins, particularly of irregular backbone portions. The aim of this work is to examine whether or not the observed chain dimensions in native proteins, expressed as the mean squares of the end-to-end distances $\langle r^2 \rangle$ or the radial distributions of end-to-end distances r, are comparable with the "unperturbed" dimensions calculated solely on the basis of a single-residue energy appropriately chosen. Throughout this work, an "unperturbed" chain means only to be devoid of explicit long-range interaction terms so that it does not imply the actual polypeptide chain to be in Θ conditions.

Procedure

The statistical analysis gives only the average properties over all the structural data (e.g., the end-to-end distance), and the differences in the sequence and the individual property of amino acids of proteins are not explicitly included in the computed properties. We, therefore, simplify the system by approximating a protein polypeptide chain as "homopolypeptide" consisting of equivalent residues of the averaged character. However, since glycine and proline possess the distinct freedoms of the internal rotations from the other amino acids (all of which are referred here to alanine-type residues), we will employ two statistics: one including glycine and proline and the other without them. Similar precaution is also taken for the regular secondary structures of α helix and β structure which obey different statistics from that of the random coil.

By using atomic coordinates of 19 proteins (see the legend of Figure 2) as the data base, the observed mean squares of end-to-end distances for a chain length n are computed as the simple average over squares of distances between C^α atoms separated by n virtual bonds along protein chains. Thus the longer the chain length, the smaller the sample size becomes. For statistics with exclusion of α helix and β structure (or glycine and proline), peptide segments that contain none of the regular secondary structures (or glycine and proline) were picked up so that the sample size diminishes significantly for large n.

For theoretical calculation of the chain dimension of a randomly coiled polypeptide, we have only considered the "unperturbed" random coils in which the long-range effects are not explicitly included as already mentioned. A value of the mean squares of end-to-end distances for an unperturbed random coil is represented as

$$\langle r^2 \rangle = C_n n l^2$$

where n is the number of virtual bonds constituting the polypeptide chain, l is the virtual bond length (=3.8 Å), and C_n is the characteristic ratio. C_n reaches a constant value with increasing n. According to Brant and Flory,[12] the unperturbed chain dimension of (homo)polypeptide can be computed based solely on the single-residue potential energy via average transformation matrix $\langle \mathbf{T} \rangle$.

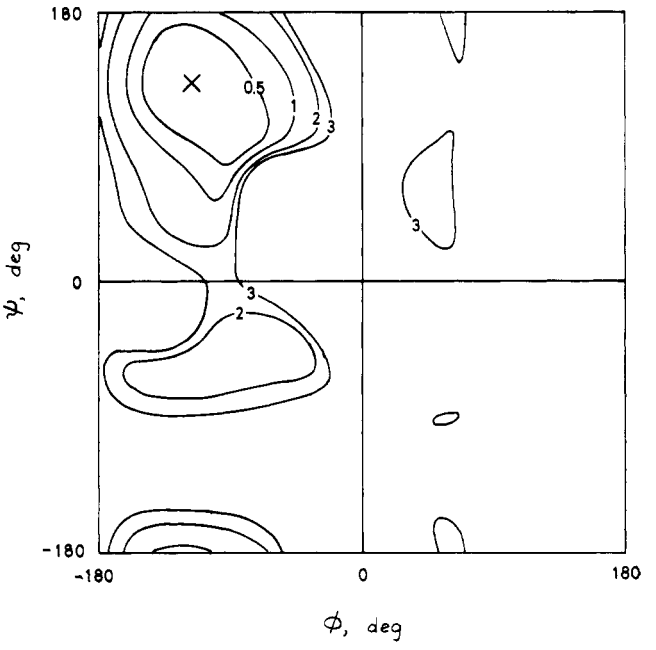

Figure 1. Atom pair potential energy map for a single residue of L-alanine. The energy contours are in kcal/mol above the global minimum (indicated with an X). The bond lengths and bond angles are those of Pauling except the bond angle $\tau(NC^\alpha C')$ fixed at 111.5°. Nonbonded energy parameters are from Oobatake and Ooi,[13] which treat the H atoms attached to C^α and C^β as the united atoms. The parameters for electrostatic energy are those of Ooi et al.[14] with the dielectric constant of 4 employed.

We considered two kinds of single-residue energy; Figure 1 illustrates the atom pair potential energy map for a structural unit of *N*-acetyl-*N'*-methyl-L-alanine, calculated as the sum of the nonbonded energy of Lennard–Jones 6–12 potential[13] and the electrostatic energy.[14] The second one is the so-called (empirical) protein energy map first presented by Pohl.[15] This energy map (Figure 2) was obtained from all the alanine-type residues in the 19 proteins by converting the observed frequency of ϕ, ψ values to the relative energy. The number of occurrences of ϕ and ψ in the ith area (20° square in size), Q_i, is related to the energy E_i as

$$Q_i/Q_0 = \exp[-(E_i - E_0)/kT]$$

where Q_0 and E_0 are the corresponding quantities at the area of the largest number of occurrence.

Radial distributions of end-to-end distances for the unperturbed chain were calculated by the Monte-Carlo method.[6-9] The statistical weight at each grid point of 30° interval on a single-residue energy map was taken into consideration for random sampling of residual conformation. Conformations of a polypeptide chain 30 virtual bonds long were generated, and at the same time conformations of shorter chains were taken from the middle part of each sample polypeptide.[6] The sample size was 3000. Radial distribution is represented with the normalized fraction, $f(r)$, i.e.,

$$\int_0^\infty f(r)\, dr = 1$$

All the computations were carried out with FACOM 230-48 at the Institute for Chemical Research of Kyoto University.

Results and Discussion

Several different features can be recognized between Figures 1 and 2. Notable among them is the mutual weight of the α_R and the extended regions: the extended region is dominant in Figure 1, whereas areas of both regions are comparable in

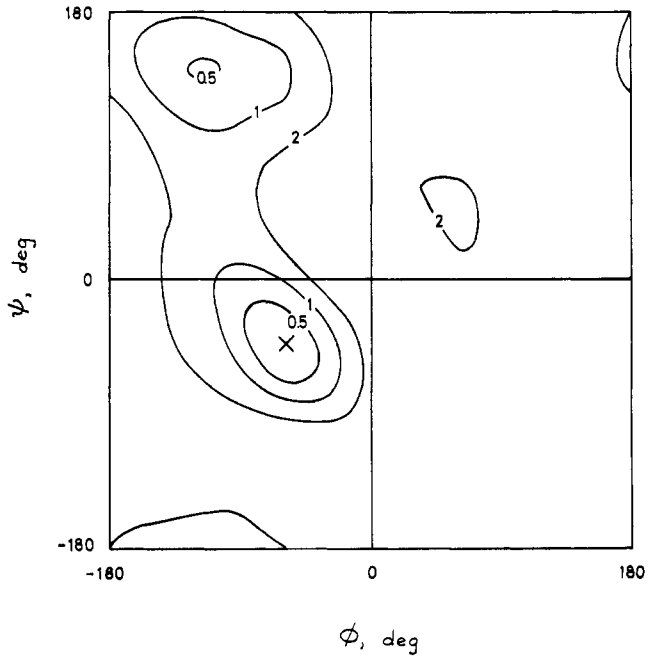

Figure 2. Empirical protein energy map constructed following Pohl[15] by use of ϕ and ψ data of 3008 alanine-type residues (excluding Gly and Pro) of 19 globular proteins listed below. The energy contours are in kcal/mol (with T = 300 °K used for the Boltzmann factor) above the lowest energy (indicated with an X). Protein data used are: glyceraldehyde 3-phosphate dehydrogenase (green subunit), apo-lactate dehydrogenase, thermolysin, carboxypeptidase A, subtilisin BPN, α-chymotrypsin, elastase (D. N. Shotton, personal communication), concanavalin A, myoglobin, staphylococcal nuclease, flavodoxin, ribonuclease S, cytochrome c_2, cytochrome b_5, hen egg-white lysozyme, carp muscle calcium-binding parvalbumin, high potential iron protein, rubredoxin, and pancreatic trypsin inhibitor. All the data except elastase were offered from Protein Databank at Brookhaven.

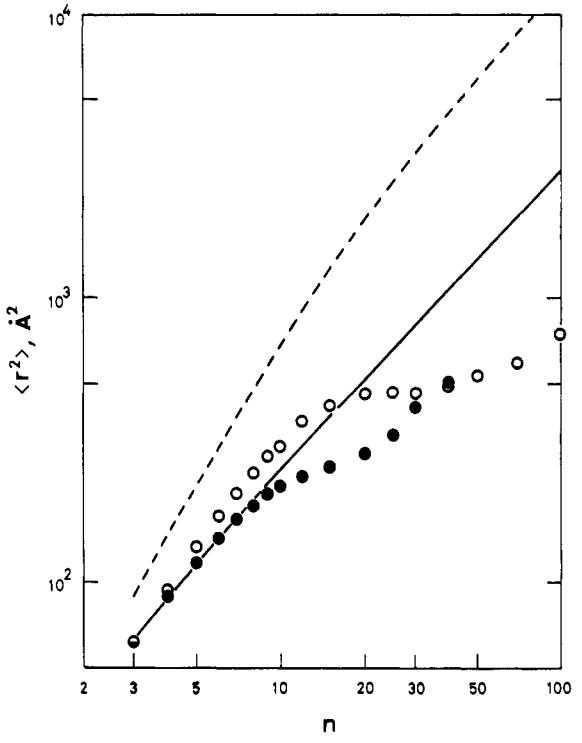

Figure 3. Logarithmic plots of the mean squares of end-to-end distances $\langle r^2 \rangle$ vs. the chain length (the number of virtual bonds) n. Observations are obtained from all the sample segments of the 19 globular proteins (○) and from the segments taken out of irregular backbone portions of the proteins (●), and theoretical calculations are for the unperturbed poly-L-alanine with use of the atom pair potential energy (---) and for the unperturbed polypeptides of alanine-type residues with the empirical protein energy (—). For theoretical calculations of $\langle r^2 \rangle$, the average transformation matrix[12] $\langle \mathbf{T} \rangle$ was computed by varying ϕ and ψ in 20° increments on a single-residue energy map.

Figure 2. The quantitative difference between these energy maps reflects more clearly in the C_n for infinitely long polypeptides; i.e., energy maps shown in Figures 1 and 2 give C_∞ = 9.46 and 1.96, respectively. The value of 9.46 derived from the atom pair potential energy is in agreement with other theoretical calculations for poly-L-alanine[6,12] and falls in the range of 8.5–10 obtained experimentally on several synthetic polypeptides.[12,16] Denatured proteins in guanidium hydrochloride solution[17] are known to have somewhat smaller values (ca. 5) than that of poly-L-alanine, because of the presence of glycine residues in proteins.[18] On the other hand, the value of 1.96 calculated by the protein energy map does not agree with any of the experimental values, but it is almost the same as the characteristic ratios of polyglycine (2.16)[19] and a polypeptide chain of freely rotatable bonds (1.93),[19] implying that the effective bond length is considerably shorter in this case compared with the other. Nevertheless the empirical protein energy map gives the chain dimension $\langle r^2 \rangle$ in agreement with the observed ones in globular proteins as shown below.

The plots of the calculations and the observations of $\langle r^2 \rangle$ against the chain length n (Figure 3) show that the observed chain dimensions for irregular backbone portions of globular proteins excluding α helices and β structures fall just on the calculated curve by the energy map of Figure 2 as long as the chain length is sufficiently short. The deviation of the observed $\langle r^2 \rangle$ for longer segments from the curve is attributed to long-range interactions: The mean squares of end-to-end distances for compact globules would be proportional to $n^{2/3}$ from a simple consideration of the volume–surface relationship. Crippen[5] discussed mainly this region of chain dimension of longer segments.

The sample segments of irregular portions in proteins used for Figure 3 contain those including glycine or proline residues, since the sample size decreases remarkably with the exclusion of these segments from the data base and therefore the statistics become unreliable. Despite the inclusion of proline and glycine residues, the comparison of the observed chain dimensions with those calculated for the "alanine-type" polypeptide may be meaningful for the characteristic ratio of polyglycine is almost the same as that obtained by the protein energy map of alanine-type residues as mentioned before. From this consideration, we can say that protein segments with short chain length up to, say, n = 8 behave like unperturbed random chains, yielding chain dimensions in good agreement with the calculated values based on the enrgy map shown in Figure 2. In turn, we note that protein segments containing the regular secondary structures deviate from the unperturbed chain dimension even for short chain length.

Further examination of whether short segments in globular proteins behave like unperturbed chains or not was made by the comparison of the observed distributions of end-to-end distances in proteins with the unperturbed chain distributions simulated by the Monte-Carlo method using the protein energy map. Figure 4 shows that the coincidence in both the distributions is fairly good for short segments of irregular protein portions while significant discrepancy is found when the regular secondary structures are included in the data set of the observations. A sharp rise at smaller r and a bulge at large r appearing in the plots including the observed secondary structures evidently originated from extension of strands of α helix and β structure, respectively. When the unperturbed chain distribution is calculated based on the atom pair energy

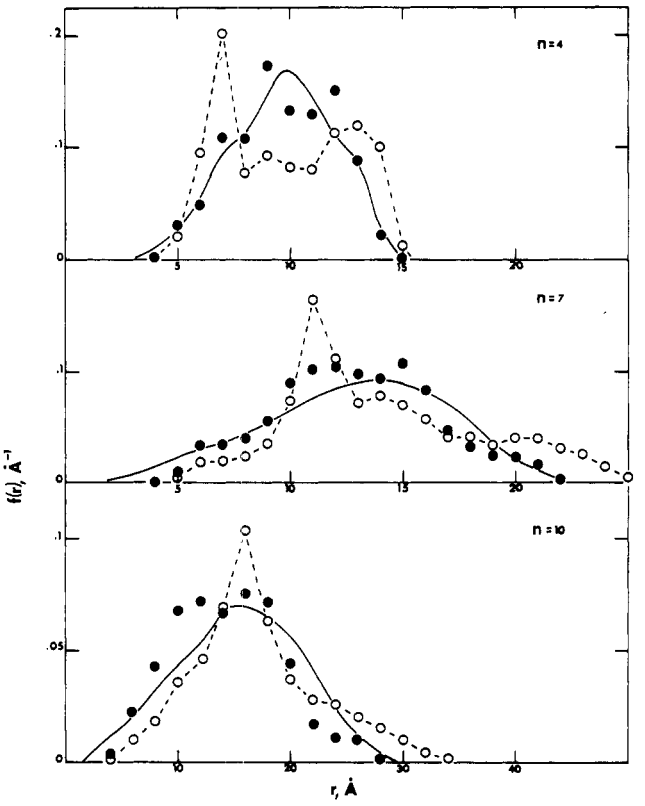

Figure 4. Normalized distributions of end-to-end distances of chains containing 4 (top), 7 (middle), and 10 virtual bonds (bottom). Open circles linked with broken lines are the observed distributions obtained from the 19 protein data including regular secondary portions, and filled circles are the corresponding distributions obtained only from irregular backbone portions. The solid curves are unperturbed chain distributions calculated from the Pohl's energy map by the Monte-Carlo method (see the text).

map, the resulting distribution completely shifts to the direction of larger r for any chain length (such examples are shown in ref 6), indicating again the entire disagreement from the observed chain distributions taken from native proteins.

These results indicate that a single residue in a globular protein, on the average, is primarily ruled by the empirical protein energy rather than the inherent (atompair) potential energy. In this respect, the protein energy should be regarded as the average potential energy for a residue in a protein, modified by interresidue interactions as well as the solvation effect. This is consistent with the demonstration by Hopfinger[20] that, while conformations of small peptides in aqueous solution are well explained with the computed energy maps, ϕ and ψ values observed in globular proteins do not necessarily distribute in accordance with the computed single-residue energy map. Suzuki and Robson[21] have also come to similar conclusions, showing that the helix nucleation parameter σ deduced from a statistical analysis of protein structures (of order of 10^{-1}) is substantially larger than experimental values (10^{-3}–10^{-4}) for synthetic polypeptides.

All these evidences may imply that the method employed here, that is, treating x-ray structural data of native proteins as a statistical ensemble for comparison with the random coil statistics, is a reasonable and effectual method to examine the average properties prevailing in the folded conformation of a native protein. We conclude that a single residue within a protein molecule obeys a mean force field which is different depending on the denatured or the native states, such as Figures 1 and 2, and that the change in effective potential energy for a single residue should take place accompanied with the folding process of a protein.

Acknowledgment. We are indebted to Dr. S. S. Zimmerman for showing us the unpublished relevant data and to Dr. M. Oobatake for the preparation of a part of the computer programs.

References and Notes

(1) P. Y. Chou and G. D. Fasman, *Biochemistry*, **13**, 211 (1974).
(2) B. Robson and E. Suzuki, *J. Mol. Biol.*, **107**, 327 (1976).
(3) S. Tanaka and H. A. Scheraga, *Macromolecules*, **9**, 945 (1976).
(4) C. Chothia, *Nature (London)*, **254**, 304 (1975).
(5) G. C. Crippen, *Biopolymers*, **16**, 2189 (1977).
(6) S. Premilat and J. Hermans, Jr., *J. Chem. Phys.*, **59**, 2602 (1973).
(7) S. Premilat and B. Maigret, *J. Chem. Phys.*, **66**, 3418 (1977).
(8) S. Tanaka and A. Nakajima, *Macromolecules*, **5**, 708 (1972).
(9) H. E. Warvari, K. K. Knaell, and R. A. Scott, *J. Chem. Phys.*, **56**, 2903 (1972).
(10) S. S. Zimmerman, L. L. Shipman, and H. A. Scheraga, *J. Phys. Chem.*, **81**, 614 (1977).
(11) R. Koyama, *J. Phys. Soc. Jpn.*, **34**, 1029 (1973).
(12) D. A. Brant and P. J. Flory, *J. Am. Chem. Soc.*, **87**, 2788, 2791 (1965).
(13) M. Oobatake and T. Ooi, *J. Theor. Biol.*, **67**, 567 (1977).
(14) T. Ooi, R. A. Scott, G. Vanderkooi, and H. A. Scheraga, *J. Chem. Phys.*, **46**, 4410 (1967).
(15) F. M. Pohl, *Nature (London), New Biol.*, **234**, 277 (1971).
(16) W. L. Mattice and J.-T. Lo, *Macromolecules*, **5**, 734 (1972).
(17) C. Tanford, *Adv. Protein Chem.*, **23**, 121 (1968).
(18) W. G. Miller, D. A. Brant, and P. J. Flory, *J. Mol. Biol.*, **23**, 67 (1967).
(19) P. J. Flory, "Statistical Mechanics of Chain Molecules", Interscience, New York, N.Y., 1969.
(20) A. J. Hopfinger, "Intermolecular Interactions and Biomolecular Organization", Wiley, New York, N.Y. 1977.
(21) E. Suzuki and B. Robson, *J. Mol. Biol.*, **107**, 357 (1976).