On the Use of Distance Constraints To Fold a Protein[1]

Hiroshi Wako and Harold A. Scheraga*

Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853.

Received November 25, 1980

ABSTRACT: A simple method is presented to assess the information that is provided by distance constraints for pairs of residues in proteins. The probability that the distance d_{ij} between the C$^\alpha$ atoms of residues i and j lies within a given range is computed for *all* $N(N-1)/2$ pairs in a molecule of N residues, and a quantity H is defined in terms of these probabilities; H is a measure of the ambiguity in the computed conformation of the molecule (consistent with the given distance constraints) and is related to the root-mean-square deviation of the computed conformation from the native one. The quantity H is used to determine the number, kind, and quality of the distance constraints required to define the conformation of a protein within given limits of error, using the 58-residue molecule bovine pancreatic trypsin inhibitor as an illustration. For example, to obtain the computed conformation with a root-mean-square deviation of less than 2 Å from the native conformation, the values of d_{ij} of more than ~80 pairs (half of them with $5 \leq |i-j| \leq 20$ and the other half with $21 \leq |i-j| \leq 57$) must be known exactly, or of more than ~150 pairs (half of them with $5 \leq |i-j| \leq 20$ and the other half with $21 \leq |i-j| \leq 57$) must be known with an error no greater than ~2 Å; alternatively, the same root-mean-square deviation of less than 2 Å from the native structure can be achieved by the computed conformation if more than ~160 pairs are chosen so that 20 Å is assigned as the lower limit for half of these d_{ij}'s (for those pairs in the native protein that are separated by ≥20 Å) and 10 Å is assigned as the upper limit for the other half of these d_{ij}'s (for those pairs in the native protein that are separated by ≤10 Å). In all of the above examples, all values of $d_{i,i+1}$ were fixed at 3.8 Å, and all values of $d_{i,i+2}$ were confined to the range 4.5–7.2 Å (the minimum and maximum possible values for a polypeptide chain). We also examined the *kind* of constraints (in terms of their distance both along the chain and through space) that are most effective to obtain a small root-mean-square deviation. For a given number of constraints, information about pairs with large $|i-j|$ or small d_{ij} is more effective in determining the conformation than is information about pairs with small $|i-j|$ or large d_{ij}. It is found, however, that information that includes both small and large $|i-j|$ or both small and large d_{ij} is the most effective.

Introduction

Concurrent with the elucidation of the amino acid sequences of insulin, ribonuclease, and lysozyme, chemical and physicochemical methodology was developed and applied to determine proximity relations between pairs of amino acid residues that could serve to define the three-dimensional structures of these proteins.[2,3] For example, using this methodology, three of the eleven carboxyl groups of ribonuclease were paired with three of the six tyrosyl groups.[3–7] With the availability of this information, together with the knowledge of the location of the four disulfide bonds[8] and the proximity of His-12, His-119, and Lys-41 in the active site,[9–13] the development of computational techniques was initiated[14] to incorporate such distance constraints in energy-minimization algorithms to try to determine the three-dimensional structure of ribonuclease.[15]

Such distance constraints restrict the conformational space of the polypeptide chain that has to be searched in a protein folding algorithm. Since distances between pairs of residues in a native protein structure can be obtained experimentally, as indicated above, and also theoretically by statistical analyses of known protein structures,[16–20] the use of such information in protein folding studies is being exploited.[18–21] The current view seems to be that, even if the number and accuracy of the distance constraints are not very large, the conformation of a protein can be predicted to some extent. It is thus of interest to ask: "What kind of distance constraints, and how many of them, are required, and how accurately must the distances be known in order to determine the conformation of a protein to any given degree of accuracy?" By "kind", we mean that the constraint depends on the distance d_{ij} through space and on the distance $|i-j|$, where i and j are sequence numbers along the chain for residues that are represented by their C$^\alpha$ atoms, and we want to know the relative effectiveness of such constraints for various values of i and j.

Havel et al.[22] considered the question of the kind of distance constraints [primarily for bovine pancreatic trypsin inhibitor (BPTI)], but not the question of the number of constraints required, or their accuracy. For example, they obtained a root-mean-square deviation of 2.9 Å between the generated and native structures when 66 distances between all pairs of Lys, Tyr, Asp, and Glu residues, and the S–S cross-links, were assigned exactly, and 7.9 Å when 200 distances corresponding to the α-helical, β-strand, and β-strand reverse-turn portions of the backbone structure were assigned exactly. This indicates that the deviation of the conformation generated under a given set of distance constraints from the native one depends not only on the number of such constraints but also on the particular pairs of residues chosen and on their accuracy, and the relation between the number and accuracy of the constraints and the deviation of the generated conformation from the native one would not be expected to be very simple. We may, however, anticipate that the average deviation of the conformations (generated under many sets of constraints having the same number and accuracy) from the native conformation would be related to the number and accuracy of the constraints. Such a relationship would give us some measure of the root-mean-square deviation that we could expect to obtain when we generate a computed conformation under a given set of distance constraints.[23] If we were to try to explore this relationship by the method of Havel et al.,[22] we would have to generate many conformations because there are many conformations consistent with a given set of constraints and, furthermore, there are many sets having the same number of constraints. Since their procedure involves optimization every time that a conformation is generated, such an effort would be too time-consuming. Therefore, a simpler algorithm is required for such an investigation.

In this paper, we consider this problem, making use of BPTI[24] as an example. In section I, we summarize the conditions under which the conformation of a protein can be determined exactly. In section II, we introduce a quantity to estimate the information contained in a given set of distance constraints and derive a relation between

0024-9297/81/2214-0961$01.25/0

this measure of information and the root-mean-square deviation of conformations (consistent with the set of distance constraints) from the native one. In section III, we use this relation and calculate the dependence of the root-mean-square deviation on the number and accuracy of the constraints; some comments are also made about the kind of distance constraints that are most effective to determine the conformation of a protein. The results are discussed in section IV.

I. Ideal Case

In this paper, we consider a protein as a system of N points representing the C^α atoms of all of its residues. Whereas N can take on any value, it is 58 for BPTI, the example for which computations are presented in sections II and III. The distance between the C^α atoms of two residues i and j is denoted by d_{ij}, and (following the elegant treatment of Crippen[25]) $\mathbf{D}$ is defined as a symmetric $N \times N$ matrix whose elements are d_{ij}; i.e., $\mathbf{D} = \{d_{ij}\}$.

Then any one of the following sets of data is sufficient to determine the conformation of a protein of N points uniquely if the data are known exactly.

(a) $3N$ Cartesian coordinates (x_i,y_i,z_i), where $i = 1, 2, ..., N$ (actually only $3N - 6$ degrees of freedom exist, the remaining 6 serving only to fix the translational and rotational positions of the whole molecule).

(b) $3N - 6$ variables distributed among $N - 1$ virtual bond lengths (C^α_i-to-C^α_{i+1} distances), $N - 2$ virtual bond angles, and $N - 3$ dihedral angles around the virtual bonds.

(c) $3N - 6$ variables distributed among $N - 1$ values of $d_{i,i+1}$, $N - 2$ values of $d_{i,i+2}$, and $N - 3$ values of $d_{i,i+3}$, *provided* that the sign of $d_{i,i+3}$ is specified [where the sign is defined according to the location of the $(i + 3)$th point with respect to the plane determined by the three points i, $i + 1$, and $i + 2$]. The three distances $d_{i,i+1}$, $d_{i,i+2}$, and $d_{i,i+3}$ determine the virtual bond length, the virtual bond angle, and the dihedral angle around the virtual bond, respectively. Since these three distances do not specify the sign of this dihedral angle, an additional $N - 4$ values of $d_{i,i+4}$ are required for this purpose, making a total of $4N - 10$ variable distances to define the conformation uniquely.[19] It is not necessary, however, to know the additional $N - 4$ values of $d_{i,i+4}$ exactly,[25] as long as they are known sufficiently accurately to specify the sign of $d_{i,i+3}$.

(d) A set of $4N - 10$ variable distances different from those specified in c ($3N - 6$ distances and an additional $N - 4$ distances to specify the sign of the distance corresponding to $d_{i,i+3}$ in c), provided that they are chosen properly. For example, if we ignore the connexity of the chain, then the N points may be numbered in any arbitrary order, rather than that determined by the amino acid sequence. In the renumbered system, $4N - 10$ variables ($d_{j,j+1}$, $d_{j,j+2}$, $d_{j,j+3}$, and $d_{j,j+4}$) are sufficient to determine the conformation uniquely. The variables in c constitute a special set in which they are numbered in the order of the amino acid sequence.[26] On the other hand, if the locations of five points i to $i + 4$ and the four distances from the $(i + 5)$th point to points $i + 1$ to $i + 4$ are known, then the location of the $(i + 5)$th point can be determined uniquely; i.e., it is not necessary to know $d_{i,i+5}$ because it can be calculated from the other distances. As in this illustration, if some distances in the set of $4N - 10$ variables are not independent, i.e., if they can be calculated from other distances in the same set, then such a set is not a proper one to determine the conformation uniquely.

In conformational studies of proteins, other sets of variables are also used, e.g., the set of backbone dihedral angles[27] ϕ_i and ψ_i, or the set of differential geometry parameters[28] curvature κ_i and torsion τ_i. Since Cartesian coordinates are convenient for distance-constraints problems,[18-20] we will not consider these other variables in this paper. The relations among the sets (ϕ_i,ψ_i), (κ_i,τ_i), and $(d_{i,i+1}, d_{i,i+2}, d_{i,i+3}, d_{i,i+4})$, however, can be derived easily by using the procedures in ref 28 and 29.

II. Method

The conformation of a protein may be determined exactly from any of the four sets of data given in section I if these data are known exactly. In actual practice, however, we cannot expect to obtain this number of distances from experiments or theoretical considerations of the kinds cited in the Introduction, nor can we expect to know these distances very accurately. Hence, the important question is: "How can the information that we can reasonably anticipate obtaining as distance constraints, by these experimental and theoretical procedures, restrict the conformation?" To answer this question, we introduce a "measure of information" and show the relation between this quantity and both the number of distance constraints and the root-mean-square deviation.

Suppose there are no constraints on the polypeptide chain except that the virtual bond lengths, $d_{i,i+1}$, are fixed at $b = 3.8$ Å. Then the molecule behaves as a freely jointed chain, for which the probability that d_{ij} lies between R and $R + dR$ is given by[30]

$$P(R)\,dR = \left(\frac{3}{2\pi kb^2}\right)^{3/2} \exp\left(-\frac{3R^2}{2kb^2}\right) 4\pi R^2\,dR \quad (1)$$

where $k = |i - j|$ and

$$\int_0^\infty P(R)\,dR = 1 \quad (2)$$

If some constraints are imposed, e.g., u_{ij} and l_{ij} as upper and lower limits, respectively, of d_{ij}, then we may define a quantity I_{ij} as

$$I_{ij} = \int_0^{l_{ij}} P(R)\,dR + \int_{u_{ij}}^\infty P(R)\,dR \quad (3)$$

Thus, if u_{ij} and l_{ij} are known, I_{ij} provides some information about d_{ij}. To put it another way

$$1 - I_{ij} = \int_{l_{ij}}^{u_{ij}} P(R)\,dR \quad (4)$$

is a measure of the ambiguity in our knowledge of d_{ij}.

If u_{ij} and l_{ij} were given for all ij pairs, the mean ambiguity of the whole system is

$$H = 1 - \frac{1}{N(N-1)/2} \sum_{i<j} I_{ij} \quad (5)$$

where $N(N - 1)/2$ is the total number of pairs. Strictly speaking, eq 1 is valid only for large k; we use it here, however, as an approximation for small k. Also, I_{ij} for a given pair is not independent of the values of I_{ij} for other pairs. We, however, are assuming that the I_{ij}'s are independent when we compute H and I_{ij} from eq 5 and 3, respectively. This dependence is, however, taken into account implicity by the values assigned to u_{ij} and l_{ij} in eq 3 and 4, since the values of these limits for a given pair do depend on the values of the limits for other pairs (see procedure b below for the calculation of u_{ij} and l_{ij}).

The quantity H is a measure of the ambiguity in the determination of the conformation of the whole protein molecule and should be related to the root-mean-square deviation of the generated conformation (consistent with the set of constraints u_{ij} and l_{ij}) from the native one. The nature of this measure of the ambiguity is discussed in the Appendix. If we can show how H is related to the root-mean-square deviation, we can assess the utility of a given set of distance constraints (i.e., estimate the root-mean-square deviation) in terms of H *without* generating the conformation under the given distance constraints, as was done by Havel et al.[22]

In order to obtain the relation between H and the root-mean-square deviation, we do not need to generate any conformations of BPTI; instead we use the mean root-mean-square deviations of the conformations already generated by Havel et al.[22] for given sets of constraints. Therefore, we first describe their procedure briefly.

(a) A set of distance constraints is given. These distance constraints are represented by upper and lower bound distance matrices $\mathbf{U} = \{u_{ij}\}$ and $\mathbf{L} = \{l_{ij}\}$, respectively, where u_{ij} and l_{ij} are upper and lower bounds for d_{ij}. The elements of $\mathbf{U}$ and $\mathbf{L}$ are assigned in various ways. Hereafter, for any set of distance constraints considered here, we set $u_{i,i+1} = l_{i,i+1} = 3.8$ Å (i.e., we fix the virtual bond length) and $u_{i,i+2} = 7.2$ Å and $l_{i,i+2} = 4.5$ Å (these being the maximum and minimum values that are possible for a *real* polypeptide chain[31]). These values ensure that the chain will be a connected one with reasonable virtual bond length and bond angles; i.e., these values restrict the virtual bond angles to a range of 143–73°, which is similar to that observed in proteins.[29] The elements of $\mathbf{U}$ beyond the first two off-diagonals (i.e., beyond $d_{i,i+2}$) were set to some specified cutoff value if the corresponding crystallographic distances were less than that cutoff value (referred to as "contacts given"). The elements of $\mathbf{L}$ beyond the first two off-diagonals were set to some specified cutoff value if the corresponding crystallographic distances were greater than that cutoff value (referred to as "noncontacts given"). In some sets of constraints, some of the elements of $\mathbf{U}$ and $\mathbf{L}$ were both set to the values of the corresponding crystallographic distance (referred to as "native distances given"). The elements of $\mathbf{U}$ and $\mathbf{L}$ beyond $d_{i,i+2}$ were sometimes obtained by also adding an auxiliary point to the chain and setting the corresponding distances from this point to each of the C^α's to the crystallographic distances between these C^α's and the center of mass (referred to as "center-of-mass distances given"). In different sets of constraints, "contacts given", "noncontacts given", "native distances given", and "center-of-mass distances given" were used either alone or in combinations of two or more.[22] The remaining elements of $\mathbf{U}$ and $\mathbf{L}$ that are not specified in the set of distance constraints are set to 40 and 5 Å, respectively; according to Havel et al.,[22] 40 Å is a reasonable maximum C^α-to-C^α distance for BPTI, and 5 Å was considered as a commonly observed minimum C^α-to-C^α distance in proteins.

(b) In a, u_{ij} and l_{ij} are assigned to an ij pair independently of other pairs. The upper and lower limits of the distances among the points are, however, related to each other according to distance geometry.[25,32] A triangle inequality is thus introduced as a necessary condition that the distances among three points must satisfy. The triangle inequality is applied[22] to every set of three points in $\mathbf{U}$ and $\mathbf{L}$. For example, if $l_{ij} < d_{ij} < u_{ij}$, $l_{jk} < d_{jk} < u_{jk}$, and $l_{ki} < d_{ki} < u_{ki}$, then

$$\begin{aligned} \max\,(l_{ij}, l_{jk} - u_{ki}, l_{ki} - u_{jk}) < d_{ij} < \min\,(u_{ij}, u_{jk} + u_{ki}) \\ \max\,(l_{jk}, l_{ki} - u_{ij}, l_{ij} - u_{ki}) < d_{jk} < \min\,(u_{jk}, u_{ki} + u_{ij}) \\ \max\,(l_{ki}, l_{ij} - u_{jk}, l_{jk} - u_{ij}) < d_{ki} < \min\,(u_{ki}, u_{ij} + u_{jk}) \end{aligned} \tag{6}$$

where max (a_1,a_2,a_3) and min (b_1,b_2) designate the maximum and minimum values, respectively, among the arguments. The elements of $\mathbf{U}$ and $\mathbf{L}$ are replaced by new values of u_{ij} and l_{ij} given by inequalities 6.

When $u_{i,i+1} = l_{i,i+1} = 3.8$ Å, $u_{i,i+2} = 7.2$ Å and $l_{i,i+2} = 4.5$ Å, and $u_{i,i+p} = 40$ Å and $l_{i,i+p} = 5$ Å (for $p = 3, 4, 5, ...$) (the latter two conditions being those given in paragraph a above), inequalities 6 lead to $u_{i,i+1} = l_{i,i+1} = 3.8$ Å, $u_{i,i+2} = 7.2$ Å, $l_{i,i+2} = 4.5$ Å, $u_{i,i+3} = 11.0$ Å, $u_{i,i+4} = 14.4$ Å, ..., and $l_{i,i+p} = 5$ Å (for $p = 3, 4, 5, ...$); $u_{i,i+p}$ (for $p \geq 5$) is given automatically by the computer program, using inequalities 6, without explicit evaluation. The above values of $u_{i,i+3}$, $l_{i,i+3}$, $u_{i,i+4}$, and $l_{i,i+4}$ are equivalent to observations on the distributions of these distances in proteins, given in Figures 3A and 4A, respectively, of ref 33.

(c) Each element of $\mathbf{D}$ is assigned a random number within the range of the corresponding elements of the new matrices $\mathbf{U}$ and $\mathbf{L}$, and the conformation is generated by a matrix method.[34] According to a theorem of Blumenthal[25,32] on the necessary and sufficient conditions for embedding N points in ordinary three-dimensional Euclidian space, if all elements of $\mathbf{D}$ are given, it is only necessary but not sufficient to satisfy the triangle inequality 6. The triangle inequality is the geometrical condition for the distances among three points, and there are higher order inequalities for the distances among four points, five points, etc. Unfortunately, it is practically impossible to obtain the matrices $\mathbf{U}$ and $\mathbf{L}$ that satisfy the necessary and sufficient conditions.[25] Therefore, some of the distances in a generated conformation may violate the bounds of $\mathbf{U}$ and $\mathbf{L}$, and it is necessary to refine the coordinates by optimizing the error function[18]

$$\sum_{i>j} A(d_{ij},u_{ij},l_{ij})$$

with respect to any d_{ij} (which violates the bounds), where

$$\begin{aligned} A(d_{ij},u_{ij},l_{ij}) &= (u_{ij}^2 - d_{ij}^2)^2 \quad \text{if } d_{ij} > u_{ij} \\ A(d_{ij},u_{ij},l_{ij}) &= 0 \quad \text{if } l_{ij} < d_{ij} < u_{ij} \\ A(d_{ij},u_{ij},l_{ij}) &= (l_{ij}^2 - d_{ij}^2)^2 \quad \text{if } d_{ij} < l_{ij} \end{aligned} \tag{7}$$

This optimization is carried out until the distances d_{ij} fall between u_{ij} and l_{ij}.

(d) Havel et al.[22] generated 10 structures for each of 18 sets of constraints for BPTI.[35] The reason for generating 10 structures is that many conformations exist for a given set of constraints, since any of the sets that they imposed did not contain sufficient information to determine the conformation uniquely. Therefore, using procedure c, they assigned 10 different sets of random numbers to $\mathbf{D}$ for each set of constraints. For each set of constraints, they computed E_x, the mean value of the root-mean-square deviation of each of the 10 structures from the X-ray crystal structure, and E_s, the mean value of the root-mean-square deviation between any two of the $\binom{10}{2}$, or 45, possible pairs of structures. They considered E_s to be some measure of the magnitude of the volume of conformational space consistent with the imposed constraints and E_x to be some measure of the average distance between the calculated and crystal structures in conformational space.

We adopted the values of E_s and E_x which Havel et al.[22] obtained by procedures a–d and also calculated H of eq 5 by numerical evaluation, using their procedures a and b and their sets of constraints to obtain all the elements of the upper and lower bound distance matrices, $\mathbf{U}$ and $\mathbf{L}$, which served as the upper and lower limits of the integrals in eq 3 and 4. [In using their procedures a and b, we optimized the values of u_{ij} and l_{ij} by means of the triangle inequality; since only the optimized values of u_{ij} and l_{ij} (and *not* the refined coordinates) are necessary for the evaluation of H, the optimization of eq 7 was not

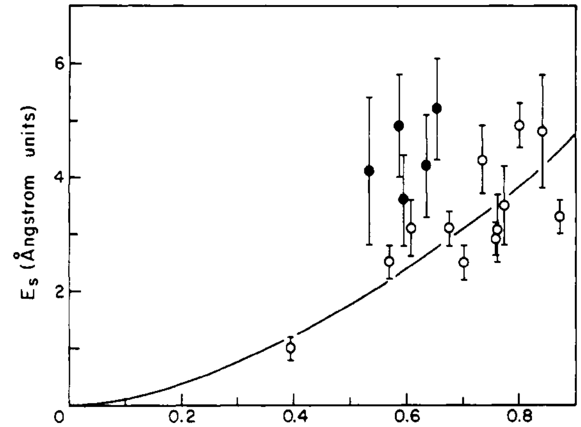

Figure 1. Relation between E_s and H. The values of E_s are those computed by Havel et al.[22] The error symbols represent one standard deviation, and the solid circles correspond to those conformations that were generated[22] under constraints that included the "center-of-mass distance". The curve (eq 8) was obtained by least-squares fitting of only the open circles (see text for details).

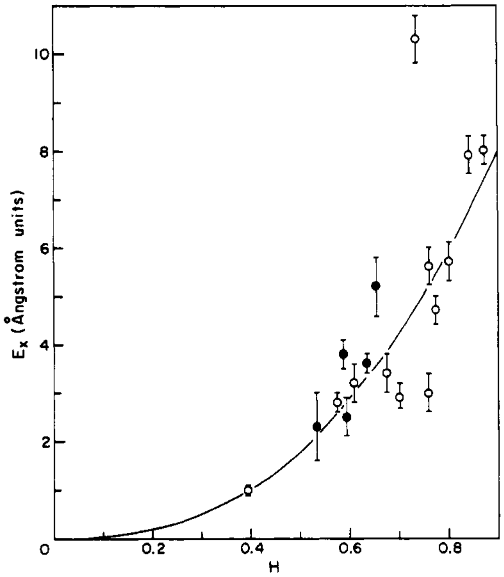

Figure 2. Relation between E_x and H. See legend of Figure 1 for further details. The curve (eq 9) was obtained by least-squares fitting of only the open circles.

carried out here.] In Figures 1 and 2, E_s and E_x, respectively, are plotted against H.

Equation 1, on which the calculation of H is based, pertains to a freely jointed chain.[30] We have, however, restricted the range of the virtual bond angles by requiring that $u_{i,i+2}$ and $l_{i,i+2}$ take on specific values. Thus, the chain is no longer a freely jointed one. Nevertheless, the form of eq 1 still applies, if b is replaced by an effective bond length b' in eq 1; i.e., the real chain is approximated by the Kuhn *equivalent* freely jointed chain[36] of bond length b'. Even though we cannot estimate the value of b', we have assumed that it is close to that of b and thus used the value of 3.8 Å for b' in computing the quantity H; this same assumption was made by Brant and Flory[37] in their calculation of the mean-square unperturbed dimensions of random polypeptide chains.

The quantity H appears to correlate well with E_x (Figure 2) but poorly with E_s (Figure 1). One of the reasons for the poor correlation between H and E_s can be understood from the following observation. In Figures 1 and 2, the solid circles represent conformations generated by Havel et al.[22] under the constraints of "center-of-mass distances given". When they incorporated the "center-of-mass distances" into their algorithm, they added an auxiliary $(N + 1)$th point (and set the distances from this point to each of the C$^\alpha$'s equal to the crystallographic distances between these C$^\alpha$'s and the center of mass) as if such a point had been an actual residue located at the center of mass. However, when only the set of distances thereby computed is used, the center of mass is not defined uniquely; i.e., this auxiliary point need not necessarily lie at the center of mass, as can be seen by the following example. In the case 5 Å < d_{ij} < 40 Å (for $i, j = 1, 2, ..., N$) and the distances $d_{i,N+1}$ given as the "center-of-mass distances" constraint (for $i = 1, ..., N$), as in run no. 12 of Havel et al.,[22] the first constraint ensures that there be no close contacts (d_{ij} > 5 Å) and that the residues not be too far from each other (d_{ij} < 40 Å); the second constraint (according to Havel et al.[22]) requires that residues C$^\alpha_1$ to C$^\alpha_N$ lie on concentric spheres with radii $d_{1,N+1}$ to $d_{N,N+1}$, respectively. This second condition, however, need not be satisfied; i.e., a knowledge of the distances $d_{1,N+1}$ to $d_{N,N+1}$ does not necessarily locate the $(N + 1)$th auxiliary point at the center of such concentric spheres. In fact, it could lie off the center and still satisfy both of the above constraints. [This is easily demonstrated by considering three points whose center of mass is at the center of three concentric spheres. Then keep two points fixed and move the third point to another position on its own sphere, but still satisfying the first condition. The (original) distances $d_{i,N+1}$ will not have changed, but the center of mass is no longer at the center of the concentric spheres.] Therefore, the incorporation of the additional $(N + 1)$th point, i.e., the use of the "center-of-mass distances" constraint, may involve some error. Another reason for the poor correlation between H and E_s may be our assumption that the distances are independent of each other. In either case, as a matter of fact, if these $(N + 1)$th points are omitted, then H correlates well with E_s also, and we will therefore not make use of the "center-of-mass distances" constraint. Inclusion of these points in Figure 2 would not disturb the good correlation between H and E_x. It is not apparent why the extra point affects the correlation with E_s, but not that with E_x.

The curves in Figures 1 and 2 were obtained by assuming a function of the form $y = ax^b$ and fitting it to the points by a least-squares procedure to optimize a and b. The results are

$$E_s = 5.64H^{1.69} \tag{8}$$

$$E_x = 10.4H^{2.53} \tag{9}$$

In this curve-fitting procedure, the solid points in Figures 1 and 2 were omitted. The correlation coefficients[38] of eq 8 and 9 are 0.88 and 0.86, respectively. Equations 8 and 9 represent the desired relations between H and the root-mean-square deviations. Having used the results of Havel et al.[22] to obtain eq 8 and 9, we make no further use of their data in the computations of section III.

III. Quantity and Quality of Distance Constraints

Since we have found a relation between H and the root-mean-square deviation, we now use it to consider how many constraints are required, and the kind and accuracy of such constraints, to determine the conformation of a protein to any given degree of accuracy.

First, in order to assess the relation between the number of constraints and the root-mean-square deviation from the native structure (i.e., in order to estimate the number

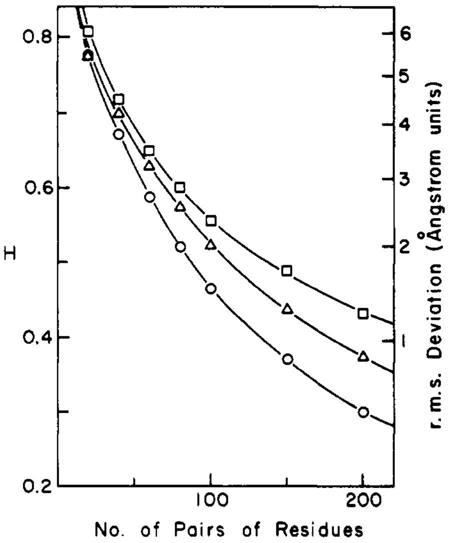

Figure 3. Relation between H (and the root-mean-square deviation E_x calculated from H by means of eq 9) and the number (n) of pairs of residues i and j. (○) case a, $n/2$ pairs with $5 \leq |i - j| \leq 20$ and $n/2$ pairs with $21 \leq |i - j| \leq 57$; (□) case b, n pairs with $5 \leq |i - j| \leq 20$; (△) case c, n pairs with $21 \leq |i - j| \leq 57$.

of constraints required to fold a protein to any given accuracy), we assign both bounds on the $N - 1$ bond lengths, $u_{i,i+1}$ and $l_{i,i+1}$, as 3.8 Å and both bounds on the $N - 2$ bond angles as $u_{i,i+2} = 7.2$ Å and $l_{i,i+2} = 4.5$ Å and choose n pairs of residues of BPTI at random, where n is less than 1653 (the maximum number of pairs). The values of both u_{ij} and l_{ij} for these n pairs are assigned the distances d^*_{ij} of the native structure. The upper and lower limits of d_{ij} (u_{ij} and l_{ij}, respectively) of the remaining $[N(N - 1)/2 - (N - 1) - n]$ pairs (allowing the values of $u_{i,i+2}$ and $l_{i,i+2}$ to change in *some* cases, to satisfy the triangle inequality) are calculated so as to satisfy the triangle inequality (eq 6), for a given set of n pairs. All of these values of u_{ij} and l_{ij}, for a given set of n pairs, correspond to the upper and lower limits, respectively, of the integrals in eq 3 and 4, so that H may be calculated from eq 5. H was calculated for 10 sets of n pairs (i.e., using a different group of n pairs for each set), and the mean value over these 10 sets was computed for the given value of n. This procedure was then repeated for different values of n, and the mean values of H are plotted against n in Figure 3 for various values of $|i - j|$. Three different sets of values of $|i - j|$ were examined, viz., (a) $n/2$ pairs with $5 \leq |i - j| \leq 20$ and $n/2$ pairs with $21 \leq |i - j| \leq 57$, (b) n pairs with $5 \leq |i - j| \leq 20$, and (c) n pairs with $21 \leq |i - j| \leq 57$. In BPTI, the maximum number of pairs in the ranges $5 \leq |i - j| \leq 20$ and $21 \leq |i - j| \leq 57$ is 728 and 703, respectively.[39]

The scale on the right-hand ordinate of Figure 3 was calculated from eq 9 Thus, for example for case a, if $n = \sim 140$, H is ~ 0.38, and the root-mean-square deviation is ~ 1 Å; i.e., ~ 140 accurately chosen distances are required to obtain a conformation with a root-mean-square deviation no greater than 1 Å, if the pairs are selected as in case a. Similarly, for this same case a, ~ 80 to ~ 140 distances are required for a root-mean-square deviation between 1 and 2 Å, ~ 60 to ~ 80 distances for a root-mean-square deviation between 2 and 3 Å, ~ 40 to ~ 60 distances for a root-mean-square deviation between 3 and 4 Å, etc. Figure 3 provides similar estimates for cases b and c.

Figure 3 provides even more information. For a given number n of distance constraints, the most effective set of constraints among the three cases examined is case a, and case c is better than case b. In other words, while information about the distances of more distant pairs along the chain (case c) is more effective than that of nearer pairs along the chain (case b), the best information is that for pairs chosen randomly from the *whole* chain. A similar conclusion was reached by Havel et al.[22] in that a knowledge of the exact distances corresponding to the α-helical, β-strand, and β-strand reverse-turn portions of the backbone structure is not sufficient to constrain the conformation of the molecule, but in conjunction with some long-range distances this information becomes more useful.

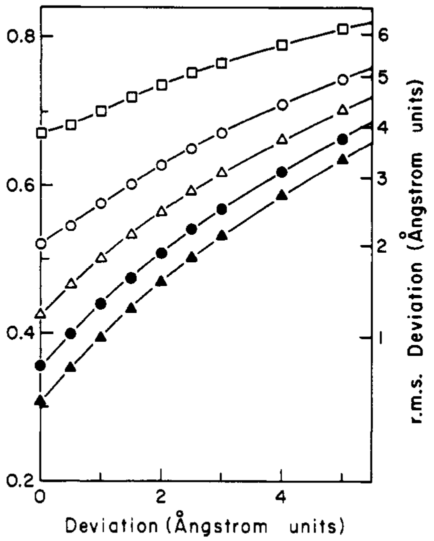

Figure 4. Relation between H (and the root-mean-square deviation E_x calculated from H by means of eq 9) and the deviation e from the native distance d^*_{ij} (i.e., the upper and lower limits are given by eq 10 and 11, respectively). The number (n) of pairs of residues i and j (for $n/2$ pairs with $5 \leq |i - j| \leq 20$ and $n/2$ pairs with $21 \leq |i - j| \leq 57$) is (□) 40, (○) 80, (△) 120, (●) 160, and (▲) 200.

To gain insight into the effect of errors in the distance constraints on the root-mean-square deviation, we examine case a in more detail. Instead of taking $u_{ij} = l_{ij} = d^*_{ij}$ for the specified n pairs, we chose instead

$$u_{ij} = d^*_{ij} + e \tag{10}$$

$$l_{ij} = d^*_{ij} - e \tag{11}$$

with u_{ij} and l_{ij} for the remaining pairs calculated with the triangle inequality, and repeated the calculations of case a. The results are shown in Figure 4. For $n = 80$ (40 pairs with $5 \leq |i - j| \leq 20$ and 40 pairs with $21 \leq |i - j| \leq 57$) as an example, the root-mean-square deviation from the native conformation increases from ~ 2 Å to ~ 3, ~ 4, and ~ 5 Å as e increases from 0 to ~ 1.5, ~ 3.5, and ~ 5 Å, respectively. For $n = 200$, the root-mean-square deviation is <1 Å if $e < 1$ Å; the root-mean-square deviation becomes 2 and 3 Å, however, if e increases to 3 and 4.5 Å, respectively. Another way of interpreting Figure 4 is to note that ~ 80 distances must be known exactly or ~ 110 or ~ 150 must be known within an error of 1 or 2 Å, respectively, to obtain a computed conformation with a root-mean-square deviation of less than 2 Å from the native one.

In order to obtain a value of $H = 0$, the condition $u_{ij} = l_{ij} = d^*_{ij}$ would have to be satisfied for all 1653 pairs, according to the definition of H in eq 4 and 5. On the other hand, a proper set of $4N - 10$, or 222, distances (selected according to criteria c or d in section I) would lead to $H = 0$ if these 222 distances were known exactly and if the elements of **U** and **L** were calculated to satisfy not only the triangle inequality 6 but also the higher order inequalities. This apparent contradiction (i.e., the need for exact knowledge of 222 rather than 1653 distances) arises from the fact that there are many sets of M distances ($M \geq 4N - 10$) which, even though known exactly, *cannot* determine the conformation uniquely because they do not constitute an independent set (see the example in set d of section I); if a *proper* set of 222 distances is known exactly, and if the triangle and higher order inequalities are applied, then the conformation is determined uniquely; i.e., $H = 0$.

Second, we examine another kind of constraint. The conformation with the smallest root-mean-square deviation ($E_x = 1.0$ Å) among those generated by Havel et al.[22] was obtained when a cutoff distance (defined in procedure a in section II) of 10 Å was used for both the "contact" and "noncontact" distances. According to eq 5, this corresponds[40] to $H = 0.395$. By selecting other cutoff distances, however, we were able to design sets of distance constraints with lower average values of H, viz., 0.300 for "contact" and "noncontact" cutoff distances of 10 and 20 Å, respectively, and 0.342 for when these were taken as 10 and 25 Å, respectively. Incidentally, there are 353, 1187, 269, and 87 pairs of distances in BPTI that are less than 10 Å, greater than 10 Å, greater than 20 Å, and greater than 25 Å, respectively. Hence, these two additional calculations indicate that a knowledge of 353 pairs whose distances are less than 10 Å, together with 269 pairs whose distances are greater than 20 Å (which gives $H = 0.300$) is better than that of 353 pairs whose distances are less than 10 Å, together with either 1187 or 87 pairs whose distances are greater than 10 or 25 Å, respectively (which gives $H = 0.395$ and 0.342, respectively).

We also examined the relation between the number of constraints and the root-mean-square deviation (with the use of the function H) for this kind of constraint, i.e., one involving a *range* of distances rather than sets of exact values that were examined in the first part of this section. The n pairs were chosen in three different ways (together with $u_{i,i+1} = l_{i,i+1} = 3.8$ Å, $u_{i,i+2} = 7.2$ Å, and $l_{i,i+2} = 4.5$ Å), with the upper limit $u_{ij} = 10$ Å for the pairs whose distances d^*_{ij} in the native structure are less than 10 Å, and with the lower limit $l_{ij} = 20$ Å for the pairs whose distances d^*_{ij} in the native structure are greater than 20 Å, viz., (a) $n/2$ pairs with $d^*_{ij} < 10$ Å and $n/2$ pairs with $d^*_{ij} > 20$ Å, (b) n pairs with $d^*_{ij} < 10$ Å, and (c) n pairs with $d^*_{ij} > 20$ Å. For each value of n, 10 values of H (for different selections among the n pairs) were computed and then averaged. The resulting values of H (and the corresponding values of the root-mean-square deviations, from eq 9) are plotted against n in Figure 5. Figure 5 differs from Figure 3 in that n pairs of distances were assigned *exactly* in Figure 3 whereas the n pairs of distances were assigned a *range* of values in Figure 5.

In case a, a knowledge of more than 350 pairs (half of them in "contact" with a cutoff distance of 10 Å and the other half in "noncontact" with a cutoff distance of 20 Å) is required to obtain a computed conformation with a root-mean-square deviation of less than 1 Å from the native one. Similarly, for this case a, 160–350 pairs are required for a root-mean-square deviation of 1–2 Å, 100–160 pairs for a root-mean-square deviation of 2–3 Å, 60–100 pairs for a root-mean-square deviation of 3–4 Å, etc.

For a given value of n, case a gives the best results, and case b is better than case c. In other words, knowledge for *both* pairs in "contact" and pairs in "noncontact" is the most effective (of the three cases considered) for predicting the conformation, and knowledge for pairs in "contact" is more effective than for pairs in "noncontact". This conclusion was also reached by Havel et al.[22] They used 10 Å (or, alternatively, 25 Å) as the cutoff distances for *both* "contact" and "noncontact" in illustrative calculations and evaluated the effectiveness of these choices of cutoff distances. We have shown, however, that more effective choices of cutoff distances exist (as measured by the value of H), e.g., 10 Å for "contact" and 20 Å for "noncontact".

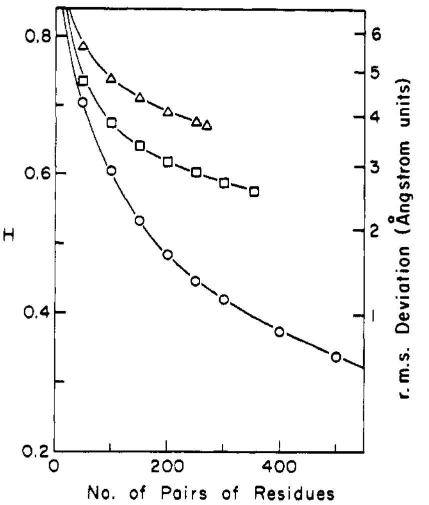

Figure 5. Relation between H (and the root-mean-square deviation E_x calculated from H by means of eq 9) and the number (n) of pairs of residues i and j. (○) case a, $n/2$ pairs with $d^*_{ij} <$ 10 Å and $n/2$ pairs with $d^*_{ij} > 20$ Å; (□) case b, n pairs with d^*_{ij} < 10 Å; (△) case c, n pairs with $d^*_{ij} > 20$ Å, where d^*_{ij} is the distance of the ij pair in the native structure.

IV. Discussion

The empirical relation between H and the root-mean-square deviation from the native conformation of BPTI presented here (eq 9) gives only a rough estimate because (1) the conformation generated under a given set of distance constraints depends not only on the number of constraints but also on which pairs of residues are in the set of constraints (i.e., *many* sets of a given number of constraints would be required), and (2) it is only a necessary but not a sufficient condition[25,32] that the triangle inequality (eq 6) be satisfied to determine the upper and lower limits of the nonspecified pairs of residues. Therefore, in discussing the relationship between the number of constraints and the root-mean-square deviation of the conformations from the native one, the results in section III should be interpreted as mean values for *many* sets of constraints (i.e., as the mean root-mean-square deviation over many sets of constraints, each of which has the same number, kind, and quality of constraints), rather than pertaining to a specific set of constraints.

On the other hand, this method can be applied as an approximation to compare the relative effectiveness of two different sets of constraints. For example, as cited in the Introduction, if we were to compare the effectiveness of knowledge of 66 distances[22] between all pairs of Lys, Tyr, Asp, and Glu residues, and the S–S cross-links, with that of 200 distances[22] corresponding to the α-helical, β-strand, and β-strand reverse-turn portions of the backbone, we might have concluded that the latter set is better than the former because it contains a larger number of constraints. If, however, we use H of eq 5 as a criterion, then the former set (with $H = 0.70$) is better than the latter set (with $H = 0.84$), and this is an alternative expression of the fact that E_x is 2.9 Å for the former and 7.9 Å for the latter.[22]

This observation also suggests that the effectiveness of the constraints depends on both the distance along the chain, $|i - j|$, and the distance in space, d_{ij}, as discussed in section III. Information about the location of disulfide bonds and hydrophobic contacts is effective because such constraints correspond to pairs with large values of $|i - j|$ and small values of d_{ij}. Further, information about pairs with large d_{ij} is much more effective when used together with information about pairs with small d_{ij}. Short-range prediction algorithms can provide information about pairs with small d_{ij}, and fluorescence–energy transfer experiments can provide information about pairs with large d_{ij}. Experiments of the type referred to in the Introduction, and statistical analyses of X-ray data on native proteins, can provide information about pairs with both small and large d_{ij}.

It is still not clear how effectively information about the distances of the residues from the center of mass constrains the conformation. For example, we could not account for the fact that the incorporation of "center-of-mass distances" by Havel et al.[22] led to discrepancies in Figure 1 but not in Figure 2. Goel et al.[19,20] also included distances of all residues from the center of mass in their algorithm but classified the residues in three categories (hydrophobic, hydrophilic, and ambivalent) and assigned mean values to these three sets of distances; using this information as well as the mean distances and the range of distances between a given residue (i.e., α carbon) and up to four nearest-neighbor α carbons and between half-cystines, they obtained conformations with root-mean-square deviations from the native structure for BPTI of around 5.3 Å. [When, on the other hand, they assigned *all* distances from the center of mass as *exact* (i.e., as given by the X-ray data), then the root-mean-square deviation from the native structure decreased to 3.08 Å.] According to our calculations, a root-mean-square deviation of 5.3 Å would be obtained if ~20 distances were known exactly (case c of Figure 3), or if 40 distances were known within an error of ~3 Å ($n = 40$ in Figure 4), or if ~15 pairs with $d^*_{ij} > 20$ Å and ~15 pairs with $d^*_{ij} < 10$ Å were assigned a range of values in "noncontact" and "contact", respectively (case c of Figure 5). Goel et al.,[19,20] however, obtained the starting conformations for their optimizations by randomly perturbing the coordinates of each α carbon by +15, -15, or 0 Å. Since there are 27 ways to induce such perturbations in the x,y,z coordinates of each α carbon, and there are 58 residues in BPTI, some of the distances between α carbons in their starting conformations were exactly the same as in the native conformation. Hence, their calculations also do not provide a test of the effectiveness of incorporation of "center-of-mass distances" in constraining the conformation.

As discussed in section I, the exact values of $3N - 6$ variables (e.g., the Cartesian coordinates or the virtual bond lengths, bond angles, and dihedral angles), or of $3N - 6$ plus $N - 4$ supplementary distances chosen properly, are sufficient to determine the conformation of a protein of N residues uniquely. For BPTI with $N = 58$, $3N - 6$ is 168. Even if 57 virtual bond lengths were known, it would be extremely difficult to obtain a proper set of 168 - 57 = 111 variables exactly by means of the kinds of experiments and theoretical considerations mentioned in the Introduction. In section III, it was shown that the exact values of about 80 distances are necessary in order to obtain a conformation with a root-mean-square deviation of ~2 Å from the native one. In actual fact, however, since we cannot know *any* values exactly, the required number of distances to determine the conformation is much greater than 80.

In this paper, distance constraints were applied to some pairs of residues and the remaining unspecified distances were restricted by application of the triangle inequality (procedure b in section II). If, however, these remaining unspecified distances are assigned a range of values rather than the exact ones, and if this range could be made narrower than that calculated with the triangle inequality (with the aid of experimental or theoretical information), then this (narrower) range of values would be more effective in restricting the conformation. Figure 4 shows that, even if there are errors in the information about d_{ij}, it is nevertheless still possible to obtain the conformation with a small root-mean-square deviation from the native one, as long as the number of known distances is large.

We may also expect that some kinds of distance constraints may be more effective than others to generate the correct conformation. For example, if we were to make use of empirical rules that enable one to predict contacts between $\alpha\cdots\alpha$, $\alpha\cdots\beta$, and $\beta\cdots\beta$ structures,[41-45] and in parallel or antiparallel arrangements,[41-45] such constraints should enhance the effectiveness of the distance geometry approach.

The use of distance constraints can be expected to provide good initial conformations for subsequent energy minimization. They can also serve as constraints in minimization and in various simulation techniques such as Monte Carlo and molecular dynamics. The method presented here provides some estimate of the number, kind, and quality of distance constraints required when protein folding is attempted by using *only* such constraints, i.e., without minimization or some simulation technique.

Appendix. Nature of the "Measure of Ambiguity"

The definition of H, adopted in eq 5 to define a measure of the ambiguity, is not the only one that could have been used. For example, following Shannon,[46] we might have used the definition

$$H = -\sum_{i<j}\log_2 (1 - I_{ij}) = -\log_2 [\prod_{i<j}(1 - I_{ij})] \quad \text{(A-1)}$$

where I_{ij} is defined in eq 3. Then, if any single distance were known exactly, i.e., if $I_{ij} = 1$ (see eq 2 and 3), H would diverge to infinity. Of course, I_{ij} would approach 1 only if the given ("exact") distance were known to an infinite number of significant figures. This unsatisfactory feature of the definition of eq A-1 arises from our assumption that the I_{ij}'s are independent of each other. At any rate, because of this feature, eq A-1 is not a satisfactory definition of H for our purposes.

While eq 5 does not have clear theoretical justification, it is defined so as to avoid the above divergence at $I_{ij} = 1$ for any single distance d_{ij} (i.e., H of eq 5 vanishes, as it should, only if $I_{ij} = 1$ for *all ij* pairs) and to be a monotonic function of I_{ij} as is eq A-1; the use of eq 5 is rationalized according to whether it satisfies reasonable assumptions (described below) about how the constraints restrict the conformation of the molecule and to whether it leads to reasonable results (good correlation of H with E_x and E_s is shown in Figures 1 and 2).

The assumptions that we make are the following: if the information that we have about a given d_{ij} leads to a large value of I_{ij} (defined in eq 3), then the conformation of the molecule is more restricted than it would be if the corresponding value of I_{ij} were small. A large value of I_{ij} would occur if, for example, l_{ij} and u_{ij} were both smaller (or both larger) than the value of R_m [$=b(2k/3)^{1/2}$] which maximizes $P(R)$ of eq 1 [which corresponds to the case of "contacts given" (or "noncontacts given") of Havel et al.[22]] or $u_{ij} - l_{ij}$ were very small (which corresponds to "native distances given"). A small value of I_{ij} would occur if $u_{ij} - l_{ij}$ were very large. Furthermore, we assume that the greater is the number of distances for which we have large values of I_{ij}, the more restricted is the conformation. If these assumptions are true, then the definition of H in eq 5 should provide a measure of the available conformational space. Since, according to Havel et all.,[22] E_x and E_s provide a measure of the volume of conformational space consistent with the imposed constraints, then H, as defined in eq 5, should correlate with E_x and E_s.

We cannot bypass the need to calculate H (by eq 5), by calculating E_x and E_s directly with the aid of $P(R_{ij})\,\mathrm{d}R_{ij}$ of eq 1, as can be shown by the following argument. Either E_x or E_s may be defined by eq A-2, viz.

$$E^2 = \frac{1}{N}\,\frac{1}{Q}\int_{l_{12}}^{u_{12}}\cdots\int_{l'_{N-1,N}}^{u'_{N-1,N}}\bar{P}(\{R_{ij}\})\bar{P}(\{R'_{ij}\}) \times \sum_{i<j}(R_{ij} - R'_{ij})^2\,\mathrm{d}R_{12}\ldots\mathrm{d}R'_{N-1,N} = \frac{1}{N}\sum_{i<j}\frac{\int_{l_{ij}}^{u_{ij}}\int_{l'_{ij}}^{u'_{ij}}P(R_{ij})P(R'_{ij})(R_{ij} - R'_{ij})^2\,\mathrm{d}R_{ij}\,\mathrm{d}R'_{ij}}{\int_{l_{ij}}^{u_{ij}}\int_{l'_{ij}}^{u'_{ij}}P(R_{ij})P(R'_{ij})\,\mathrm{d}R_{ij}\,\mathrm{d}R'_{ij}} \qquad \text{(A-2)}$$

where the normalizing factor Q and the probability $\bar{P}(\{R_{ij}\})$ for the chain to take on a given conformation $\{R_{ij}\}$ (if the residues are assumed to be independent of each other) are given by

$$Q = \int_{l_{12}}^{u_{12}}\cdots\int_{l'_{N-1,N}}^{u'_{N-1,N}}\bar{P}(\{R_{ij}\})\bar{P}(\{R'_{ij}\})\,\mathrm{d}R_{12}\ldots\mathrm{d}R'_{N-1,N} \qquad \text{(A-3)}$$

and

$$\bar{P}(\{R_{ij}\}) = \prod_{i<j}P(R_{ij}) \qquad \text{(A-4)}$$

where $P(R_{ij})$ is given by eq 1 and R_{ij} and R'_{ij} refer to two structures that are being compared. The average of each term in eq A-2, however, is taken only over the constrained conformational space [i.e. normalized by $\int_{l_{ij}}^{u_{ij}}\int_{l'_{ij}}^{u'_{ij}}P(R_{ij})\cdot P(R'_{ij})\,\mathrm{d}R_{ij}\,\mathrm{d}R'_{ij}$] and does not inform us what the volume of the constrained conformational space is, as compared with the whole conformational space [i.e., the ratio of $\int_{l_{ij}}^{u_{ij}}\int_{l'_{ij}}^{u'_{ij}}P(R_{ij})P(R'_{ij})\,\mathrm{d}R_{ij}\,\mathrm{d}R'_{ij}$ to $\int_0^\infty\int_0^\infty P(R_{ij})P(R'_{ij})\,\mathrm{d}R_{ij}\,\mathrm{d}R'_{ij}$]. Hence, we do not calculate E_x and E_s in this manner.

In summary, in comparison with the calculations of eq A-2 to A-4, the use of H, as defined in eq 5, is simple and reasonable in the sense that it satisfies the assumptions stated above and that E_x and E_s correlate well with H.

References and Notes

(1) This work was supported by research grants from the National Science Foundation (PCM79-20279) and from the National Institute of General Medical Sciences of the National Institutes of Health, U.S. Public Health Service (GM-14312).

(2) Scheraga, H. A. "Protein Structure"; Academic Press: New York, 1961; pp 241–83.

(3) Scheraga, H. A. In "Current Topics in Biochemistry—1973"; Anfinsen, C. B., Schechter, A. N., Eds.; Academic Press: New York, 1974; pp 1–3.

(4) Scheraga, H. A. *Fed. Proc., Fed. Am. Soc. Exp. Biol.* **1967**, *26*, 1380.

(5) Considering that there are $6\binom{6}{3}\binom{11}{3}$ or 19 800 ways to pair 3 of 6 tyrosyls with 3 of 11 carboxyls, this specific pairing (which is consistent with the subsequently determined X-ray structure of ribonuclease[6,7]) represented the unique result of a long series of chemical and physicochemical studies.[4]

(6) Kartha, G.; Bello, J.; Harker, D. *Nature (London)* **1967**, *213*, 862.

(7) Wyckoff, H. W.; Tsernoglou, D.; Hanson, A. W.; Knox, J. R.; Lee, B.; Richards, F. M. *J. Biol. Chem.* **1970**, *245*, 305.

(8) Spackman, D. H.; Stein, W. H.; Moore, S. *J. Biol. Chem.* **1960**, *235*, 648.

(9) Barnard, E. A.; Stein, W. D. *J. Mol. Biol.* **1959**, *1*, 339, 350.

(10) Gundlach, H. G.; Stein, W. H.; Moore, S. *J. Biol. Chem.* **1959**, *234*, 1754, 1761.

(11) Heinrikson, R. L.; Stein, W. H.; Crestfield, A. M.; Moore, S. *J. Biol. Chem.* **1965**, *240*, 2921.

(12) Heinrikson, R. L. *J. Biol. Chem.* **1966**, *241*, 1393.

(13) Hirs, C. H. W.; Halmann, M.; Kycia, J. H. In "Biological Structure and Function"; Goodwin, T. W., Lindberg, O., Eds.; Academic Press: New York, 1961; Vol. 1, p 41.

(14) Nêmethy, G.; Scheraga, H. A. *Biopolymers* **1965**, *3*, 155.

(15) Scheraga, H. A. In "Protein Folding"; Jaenicke, R., Ed.; Elsevier: Amsterdam, 1980; p 261.

(16) Tanaka, S.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 945.

(17) Crippen, G. M. *Biopolymers* **1977**, *16*, 2189.

(18) Kuntz, I. D.; Crippen, G. M.; Kollman, P. A. *Biopolymers* **1979**, *18*, 939.

(19) Ycas, M.; Goel, N. S.; Jacobsen, J. W. *J. Theor. Biol.* **1978**, *72*, 443.

(20) Goel, N. S.; Ycas, M. *J. Theor. Biol.* **1979**, *77*, 253.

(21) Cohen, F. E.; Sternberg, M. J. E. *J. Mol. Biol.* **1980**, *137*, 9.

(22) Havel, T. F.; Crippen, G. M.; Kuntz, I. D. *Biopolymers* **1979**, *18*, 73.

(23) It should be noted that we are *not* interested here in the average of a set of generated computed conformations as a representation of the native structure of the protein. Instead, as stated in the text, we use this average only to provide some measure of the expected root-mean-square deviation for any *given* generated conformation.

(24) The Cartesian coordinates were obtained from the Protein Data Bank at Brookhaven National Laboratory. See: Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535.

(25) Crippen, G. M. *J. Comput. Phys.* **1977**, *24*, 96.

(26) For example, assume that there are N points in space, numbered sequentially from 1 to N. In set c, $4N - 10$ distances suffice to locate these N points, e.g., d_{12}, d_{23}, d_{34}, ..., $d_{N-1,N}$, d_{13}, d_{24}, ..., $d_{N-2,N}$, d_{14}, d_{25}, ..., $d_{N-3,N}$, d_{15}, ..., $d_{N-4,N}$. If we change the numbering of these points, ignoring their original connexity, as in set d, we may assign the numbers 1, 2, 3, 4, 5, ... to those that were originally called, for example, 10, 4, 28, 15, 21, ... Thus, e.g., the pairs 1–2, 2–3, 3–4, ... in the renumbered system correspond to the pairs 10–4, 4–28, 28–15, ..., respectively, in the original system. Hence, using the original numbering as subscripts, the following $4N - 10$ distances also suffice to determine the location of these N points: $d_{10,4}$, $d_{4,28}$, $d_{28,15}$, ..., $d_{10,28}$, $d_{4,15}$, ..., $d_{10,15}$, $d_{4,21}$, ..., $d_{10,21}$, ... Even though the $4N - 10$ distances of this set are chosen irrespective of the connexity of the chain, this is a set of independent distances, and therefore as good a set as the original one for determining the conformation uniquely. The above procedure for renumbering the points is only one example to illustrate that there exist at least $N!$ proper sets of $4N - 10$ independent distances [$N!$ is the number of ways to number (or permute) N points].

(27) IUPAC–IUB Commission on Biochemical Nomenclature *Biochemistry* **1970**, *9*, 3471.

(28) Rackovsky, S.; Scheraga, H. A. *Macromolecules* **1978**, *11*, 1168. *Ibid.* **1980**, *13*, 1440.

(29) Nishikawa, K.; Momany, F. A.; Scheraga, H. A. *Macromolecules* **1974**, *7*, 797.

(30) Flory, P. J. "Principles of Polymer Chemistry"; Cornell University Press: Ithaca, N.Y. 1953; p 407.

(31) Momany, F. A.; McGuire, R. F.; Burgess, A. W.; Scheraga, H. A. *J. Phys. Chem.* **1975**, *79*, 2361.

(32) Blumenthal, L. M. "Theory and Applications of Distance Geometry", 2nd ed.; Chelsea: New York, 1970; pp 98–105.

(33) Isogai, Y.; Nêmethy, G.; Rackovsky, S.; Leach, S. J.; Scheraga, H. A. *Biopolymers* **1980**, *19*, 1183.

(34) Crippen, G. M.; Havel, T. F. *Acta Crystallogr., Sect. A* **1978**, *A34*, 282.

(35) In our computations here, we omitted their set number 5 because the hydrophilic residues for this set of constraints were not sufficiently well-defined. The hydrophobic residues in their other sets of constraints were defined in ref 18.

(36) Reference 30, pp 411–13.

(37) Brant, D. A.; Flory, P. J. *J. Am. Chem. Soc.* **1965**, *87*, 2791.

(38) The correlation coefficients were calculated by linear regression analysis for eq 8 and 9 in the form $\log y = b \log x + \log a$, i.e., as $[\langle \log y \log x\rangle - \langle \log y\rangle\langle \log x\rangle]/[\langle(\log y)^2\rangle - \langle \log y\rangle^2]^{1/2}[\langle(\log x)^2\rangle - \langle \log x\rangle^2]^{1/2}$, where $\langle\cdots\rangle$ stands for the mean value of a given quantity, y corresponds to E_s or E_x, and x corresponds to H.

(39) The values 20 (and 21) were chosen because the *total* number of pairs in each interval, $5 \le |i - j| \le 20$ and $21 \le |i - j| \le 57$ are approximately equal, viz., 728 and 703, respectively. There are, in addition, 57, 56, 55, and 54 pairs of distances $d_{i,i+1}$, $d_{i,i+2}$, $d_{i,i+3}$, and $d_{i,i+4}$, respectively. The values of u_{ij} and l_{ij} for these distances were fixed (see procedure b of section II). Then the

values of $u_{i,i+1}$ and $l_{i,i+1}$ were kept at 3.8 Å, and those for $d_{i,i+2}$, $d_{i,i+3}$, and $d_{i,i+4}$ were adjusted by the triangle inequality. In general, the values of u_{ij} and l_{ij} for small $|i-j|$ have little effect in constraining the conformation of the whole protein.

(40) A cutoff distance of 10 Å for both the "contact" and "noncontact" distances means that u_{ij} = 10 Å and l_{ij} = 5 Å if d^*_{ij} < 10 Å, and u_{ij} = 40 Å and l_{ij} = 10 Å if d^*_{ij} > 10 Å. The triangle inequalities (6) are then applied to every set of three points to modify all of the u_{ij}'s and l_{ij}'s, except $u_{i,i+1}$ and $l_{i,i+1}$, which are then used to calculate H.

(41) Sternberg, M. J. E.; Thornton, J. M. *J. Mol. Biol.* **1976**, *105*, 367. *Ibid.* **1977**, *110*, 269, 285.

(42) Richardson, J. S. *Nature (London)* **1977**, *268*, 495.

(43) Chothia, C.; Levitt, M.; Richardson, D. *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 4130.

(44) Richmond, T. J.; Richards, F. M. *J. Mol. Biol.* **1978**, *119*, 537.

(45) Cohen, F. E.; Sternberg, M. J. E.; Taylor, W. R. *Nature (London)* **1980**, *285*, 378.

(46) Shannon, C. E. *Bell Syst. Tech. J.* **1948**, *27*, 379, 623.

NMR and ESR Study of the Conformations and Dynamical Properties of Poly(L-lysine) in Aqueous Solutions

B. Perly,* Y. Chevalier, and C. Chachaty

Département de Physico-Chimie, Centre d'Etudes Nucléaires de Saclay, 91191 Gif-sur-Yvette Cedex, France. *Received August 13, 1980*

ABSTRACT: The conformations and dynamical behavior of poly(L-lysine) (PLL) in aqueous solutions have been investigated by ^{1}H and ^{13}C NMR as well as by ESR on the end-chain spin-labeled polymer. The ESR allowed the motion of the macromolecular chain to be studied up to pH 13, showing that the random coil → α-helix transition at pH 11 gives rise to a twofold increase in the correlation time, with evidence of an anisotropic reorientation. In the random coil state at pH 7, where the segmental motion of the backbone is quasi-isotropic, the correlation time given by ESR is compared to that obtained by the relaxation of the methyl protons of the reduced Tempo radical residue and of the α carbons. The different methods yield an activation energy of 6.5 kcal mol^{-1} for this motion whereas the frequency dependence of the C$_\alpha$ relaxation may be interpreted by a Cole–Cole distribution of correlation times with a width parameter γ = 0.7. The rotational isomerism and temperature dependences of interconversion rates of the aminobutyl side chains have been analyzed from the proton vicinal couplings and the ^{13}C and ^{1}H relaxation at different frequencies, assuming that the methylene groups undergo 120° jumps among three sites, two of them being equiprobable. These two kinds of information concur to show that the PLL side chains are less flexible than a hydrocarbon chain of same length, possibly because of the hydration of the NH$_3$$^+$ terminal group.

Introduction

Among homopolypeptides, which may be considered as the simplest models for natural proteins, poly(L-lysine) (PLL) has been subjected to a great deal of study on its conformational properties as well as its biological activity.[1]

In aqueous solution, poly(L-lysine) is known to exist in several forms, namely, random coil, α helix, and β sheets, depending upon pH and temperature. The random coil → α-helix transition which occurs around pH 11 has been investigated by several NMR techniques,[2] in particular by ^{1}H chemical shifts[3] and ^{13}C longitudinal relaxation,[4] the latter method giving semiquantitative information on the segmental mobility of the polymer. More recently, poly(L-lysine) in the α-helix form was taken as a model in a theoretical study of the motion of an alkyl chain attached to a rigid rod undergoing an anisotropic overall motion.[5]

The present work deals mainly with the dynamical behavior and the conformational properties of poly(L-lysine) in the random coil state by ^{1}H and ^{13}C NMR and relaxation, i.e., below pH (or pD) 10, where well-resolved spectra may be obtained. Special attention has been paid to the relationship between the nuclear relaxation data, the proton vicinal coupling constants, and the rotational isomerism about each of the C–C bonds of the aminobutyl side chains.

As a complement to the NMR studies, ESR experiments on the spin-labeled polymer provide a straightforward determination of the segmental motion of the main chain in both random coil and α-helix structures. A direct comparison with proton relaxation data has been provided by a diamagnetic analogue of the spin label.

Experimental Section

Materials. Poly(L-lysine) has been prepared by polymerization of L-lysine, the ϵ-amino group being protected by trifluoroacetylation. This procedure was preferred to the original one of Fasman et al.[6] because the group must be removable under mild conditions, particularly in the case of a spin-labeled polymer. N^ϵ-(Trifluoroacetyl)-L-lysine was prepared from L-lysine and *S*-ethyl trifluorothioacetate according to the procedure of Calvin et al.[7] Conversion to N^ϵ-(trifluoroacetyl)-L-lysine *N*-carboxyanhydride (N^ϵ-TFA-L-Lys-NCA) was performed by treatment with 4 M phosgene solution in tetrahydrofuran.[8] Prior to use NCA was recrystallized from ethyl acetate/petroleum ether. N^ϵ-TFA-L-Lys-NCA [2.68 g (10^{-2} mol)] was dissolved in 25 mL of anhydrous dimethylformamide. After addition of 10 mg (10^{-4} mol) of *n*-hexylamine (monomer/initiator ratio = 100), polymerization was allowed to proceed at room temperature under continuous stirring for 2 days. Precipitation in 100 mL of water yielded 2.0 g (89%) of poly[N^ϵ-(trifluoroacetyl)-L-lysine]. The trifluoroacetyl group was removed by dissolving 0.34 g of poly[N^ϵ-(trifluoroacetyl)-L-lysine] into 7.5 mL of a 1 M piperidine solution in methanol. After 2 h, 5 mL of 1 M aqueous piperidine was added dropwise under stirring to the latter solution. After 2 days, the resulting clear solution was dialyzed for 5 days against circulating distilled water at 5 °C and then against a 10^{-3} M HCl aqueous solution for 2 days. Finally the solution was freeze-dried, yielding 205 mg (80%) of poly(L-lysine) hydrochloride as a white fibrous material.

The spin-labeled poly(L-lysine) (Tempo-PLL) was synthesized following the same procedure as reported above, replacing *n*-hexylamine by 17 mg (10^{-4} mol) of 4-amino-2,2,6,6-tetramethylpiperidinyl-*N*-oxy (Tempo) as initiator. The diamagnetic