An Approach to the Multiple-Minimum Problem in Protein Folding, Involving a Long-Range Geometrical Restriction and Short-, Medium-, and Long-Range Interactions[1]

H. Meirovitch[2a] and H. A. Scheraga*[2b]

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

Received January 5, 1981

ABSTRACT: Because of the extremely large number of minima in the multidimensional conformational energy hyperspace of a protein, it is very improbable that the native structure can be attained by minimizing the energy of a conformation that does not have the same *long-range* structural features as the native molecule. To attempt to circumvent this difficulty, we divide the conformational space of a protein [bovine pancreatic trypsin inhibitor (BPTI), in the example considered here] into a relatively small number of classes according to the spatial geometric arrangements of the loops (SGAL) defined by the disulfide bonds (three, in the case of BPTI). To reduce computer time, we suggest that each of these classes (possibly ~20 in number for BPTI) be represented by at least one conformation chosen from it at random and that the conformational energy of each of these selected structures be minimized. We assume that one of the conformations selected from the class containing the native structure would emerge as the one of lowest energy. The selected conformations are defined *initially* with the help of a space-filling model (SFM) of BPTI. The bond lengths, bond angles, and energy are defined by the computer program ECEPP (which takes all backbone and side-chain atoms, including nonpolar hydrogens, into account), and the energy is minimized by alternating between two different procedures. To test this approach in this initial study, we minimized the energy of only three starting conformations: (a) The "native" structure, (b) a conformation having the same SGAL as the native structure, but differing from it in other structural aspects (e.g., it did not have the C-terminal α-helix), and (c) a conformation which did not have the correct SGAL. The energy of (a) was lowered from 10^6 to -179 kcal/mol (primarily because of relief of a few steric overlaps) without significant changes in the dihedral angles, indicating that the native structure is one of minimum energy in the ECEPP representation. The energy of (b) was lowered from 10^{10} to +94 kcal/mol (with a root-mean-square deviation of 5.9 Å) but, even though the chain had the correct SGAL, the structure differed from the native one in many respects (e.g., it still did not have the C-terminal α-helix, and it had a larger radius of gyration). From this result, we conclude that energy minimization of a compact conformation, by itself, does not lead to drastic changes in structure. The SFM and the correct SGAL greatly restrict the conformation, from both a short- and a long-range point of view, respectively, but while these restrictions are necessary they are not sufficient to fold the chain correctly without the inclusion of additional short-, medium-, and long-range information. The minimized energy of (c) is +168 kcal/mol, which is higher than that obtained from (b). This is in accord with our assumption that one of the structures with the correct SGAL would lead to the lowest energy. The efficiency of the two minimization procedures, the technical problems in the use of the SFM, and the directions for future work are discussed.

Introduction

A currently accepted assumption, based on experiments of Anfinsen and co-workers,[3-6] is that the native structure of a globular protein corresponds to the global minimum of its free energy. A statistical mechanical treatment of a *detailed* model of even a small protein such as bovine pancreatic trypsin inhibitor (BPTI) in a solvent or in vacuum (by energy minimization,[7] Monte Carlo,[8] or molecular dynamics procedures[9]) is, however, beyond the scope of present computers. The essential difficulty is the multiple-minimum problem, i.e., the existence of an extremely large number of local minima in the multidimensional conformational energy hyperspace of the protein. In the course of minimization of the energy, the system is trapped in a minimum close to the starting conformation, which in general is not the global minimum but a local one. Therefore, if minimization is started from a conformation chosen at random, the probability of reaching the global minimum is essentially zero.[10-12]

In view of this difficulty, a two-stage strategy has been adopted to attempt to reach the global minimum.[13] In the first stage, approximate methods are used to obtain structures which, in some aspects, resemble the native one. In the second stage, these structures are refined by means of rigorous energy minimization. In most cases, however, the second stage has not been carried to completion because of the large amount of computer time required.[7] Most of these approximate methods are summarized in ref 14; it is, however, of interest to discuss some of them here in order to demonstrate the need for developing new procedures such as the one to be presented in this paper.

In one method, highly simplified models for the protein (which are assumed to have relatively smooth energy surfaces) are developed and the energy is minimized.[15] This approach has been used by Levitt and Warshel[16,17] and by Kuntz et al.[18] They started minimization from open chains and obtained compact structures. The results, however, were unsatisfactory, and the multiple-minimum problem was not surmounted in these models. In another approach, short-range interaction models are used. For example, the probabilities for an amino acid to be located in the helix state, the extended state, etc. are derived from X-ray structures of many proteins (some of these procedures are given in ref 19-23 and others are reviewed in ref 14 and 19). Burgess and Scheraga[7] showed that this *short-range* procedure is also unsatisfactory, even if the prediction is a perfect one, since the ranges of the states in the (ϕ, ψ) maps are relatively large and a given sequence of states can therefore lead to substantially different structures.[24,25] In order to decrease the conformational range of the chain substantially, information about *long-range* contacts must also be incorporated. Such information has been used by Tanaka and Scheraga[26,27] in a simple model for a protein in water; the results were unsatisfactory, however, probably because of an oversimplified treatment of the chain and because a sufficient number of long-range constraints was not included in this

0024-9297/81/2214-1250$01.25/0

procedure (the question of the number, kind, and quality of constraints required has been discussed elsewhere[28]). One type of long-range constraint, an artificial attractive potential to form disulfide bonds between appropriately paired half-cystine residues, is included in the detailed energy function ECEPP,[29] which also treats *all* pair interatomic interactions (short-, medium-, and long-range). Knowledge of the location of the disulfide bonds, however, is not sufficient to define the protein uniquely (without carrying out energy minimization[28] and, possibly, even when energy minimization is carried out); in any event, incorporation of the known locations of disulfide bonds does not constitute enough information to surmount the multiple-minimum problem.

A purely geometric approach (in contrast to the energetic ones discussed above) has been suggested recently by Crippen,[30] by Kuntz et al.,[31] by Havel et al.,[32] by Ycas et al.,[33] and by Goel and Ycas.[34] In these procedures, all sources of information (including statistical data from X-ray structures of proteins) for determining restrictions on short-range as well as medium- and long-range distances between the atoms of the protein are used. This information is incorporated in an "objective" function which is optimized with respect to the atomic coordinates. The chain is represented *only* by point-mass C^α atoms, and efficient optimization procedures are used. These simplifications smooth out the surface of the objective function and speed up the rate of convergence; thus, this approach is useful to study the effect of the imposition of various distance constraints on the extent of folding.

In the present paper we develop a new approach to the problem of the folding of globular proteins based on energetic and geometric criteria. In the first stage, short-, medium-, and long-range interactions and a long-range geometric criterion are taken into account in order to define a set of starting conformations which is shown to be sufficiently small to be computationally feasible but sufficiently large to have a reasonable chance of including a native-like conformation. In the second stage, the energies of these conformations are minimized. Our method is applied to BPTI. The first step is to build a structure (having the amino acid sequence of this molecule) by means of a space-filling model (SFM) and to connect the three pairs of half-cystines forming the disulfide bonds (for details, see the Methods section). Paper dials are attached to the faces of certain atoms to enable us to determine the set of dihedral angles (ϕ, ψ, and χ's) corresponding to a given three-dimensional structure of the SFM. Together with standard geometry[29] (bond lengths and bond angles), these dihedral angles constitute the complete data required for defining the initial structure with the help of a computer.

Even though the model takes into account the short-, medium-, and long-range excluded volume effect and has the correct disulfide bonds, the number of SFM conformations is still very large. Further, these disulfide-bonded conformations are generally relatively compact and, because of the multiple-minimum problem, it is unlikely that the overall long-range structure can be drastically altered by energy minimization. Therefore, we introduce an additional *long-range* restriction to limit the magnitude of the computational problem and increase the chance of reaching the native structure by energy minimization. A convenient way to describe the long-range structure of a disulfide-bonded protein is to specify the spatial geometric arrangement of the loops (SGAL) formed by its disulfide bonds. These loops can penetrate or be threaded into each other in various ways which thereby define classes of structures (see Methods section). Since a conformation would not be expected to shift from one class to another by means of energy minimization, it is reasonable to expect that an adequate representation of the conformational space of a disulfide-bonded protein molecule will be obtained by first sampling one conformation from each class and minimizing its energy; this is the only (long range) restriction used to choose a representative conformation, at random, from each class. Since the native structure belongs to one of these classes, we expect that the conformation representing this class will lead to the "best" structure, i.e., the one with lowest energy. This expectation is based on the assumptions that the native structure corresponds to the global minimum of the energy and that the closer a conformation is (structurally) to the native one, the lower is its energy. These assumptions might fail for large loops, e.g., where much conformational freedom still exists for the class that contains the native structure. In this case, more than one conformation is required to represent a class; in fact, one should choose more than one representative of a class for larger loops. This approach is feasible, of course, as long as the number of classes is small. The procedure described thus far (involving a *detailed* model of a protein) does not use information from the known X-ray structures of proteins but rather is based primarily on energetic considerations. It should be pointed out, however, that empirical information can be incorporated into it as well (see Methods section). Because of the large amount of computer time required for energy minimization, we do not carry out the complete procedure in this *initial* study but rather construct and minimize the energy of only three starting conformations. Our main purpose here is primarily to test the procedure and, secondarily, to examine whether the imposition of the correct SGAL constitutes a sufficient constraint to fold the chain correctly by means of energy minimization.

In the Methods section, we describe our approach in detail and in the Results and Discussion section, the results of the work are presented and the conclusions for further development of the method are discussed.

Methods

In the first stage, a Corey–Pauling–Koltun (CPK) SFM of the backbone and side chains of BPTI was constructed, and the half-cystines were connected to form the 5–55, 14–38, and 30–51 disulfide bonds. All amino acids were placed in the L configuration. In order to preserve the planarity of the peptide group, adjacent C′ and N atoms were cemented together in the trans conformation, i.e., with $\omega = 180°$. A paper dial was attached to the face of one of the atoms for each bond about which rotational freedom was allowed, and a scratch mark was inscribed on the face of the opposite atom as described by Yankeelov et al.[35] It was thus possible to read the values of the dihedral angles (ϕ_i, ψ_i, $\chi_i{}^j$) of each residue for any conformation of the chain with a precision of ±5° (see ref 35). The criterion for placing the SFM in a specific conformation is described below.

Such a model is clearly a much more detailed one than that used in other methods.[15-18,26,31-34] In addition to the inclusion of every atom (including all hydrogens), the planarity of the peptide group is preserved, and short-, medium-, and long-range steric overlaps are *automatically* avoided because of the hard-sphere space-filling character of the CPK models. Also, some long-range contacts (the disulfide bonds) are automatically preserved in the SFM.[36,37] Another advantage of the SFM is the convenience in defining and analyzing structures with different SGAL's.

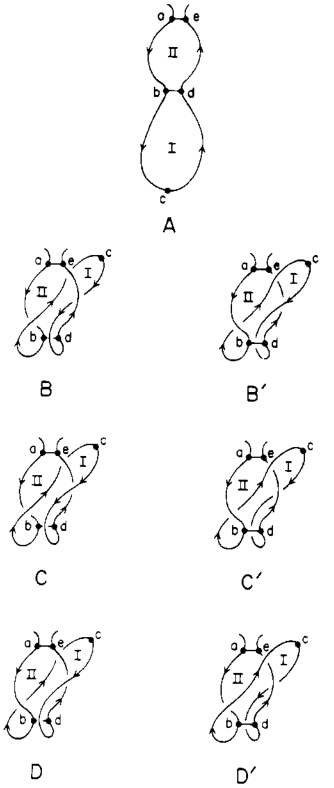

Figure 1. Schematic representation of the seven SGAL's in a protein molecule having two disulfide bonds. (A) Loops I and II are separated; (B, B′) two different ways to thread loop I through loop II (from two different sides of loop II); (C, C′) two different ways to intertwine loops I and II; (D, D′) two different ways to thread loop II through loop I (from two different sides of loop I).

The flexibility of the chain is considerably reduced by forming the disulfide bonds, but an extremely large number of conformations is still possible (see the discussion by Nêmethy and Scheraga[38]). Hence (see Introduction), random sampling of starting conformations from the whole conformational space of the disulfide-bonded protein molecule will probably lead to conformations that are structurally different from the native one; i.e., such randomly chosen starting conformations are inadequate to reach the native structure by energy minimization. We therefore suggest that the conformational space by divided into a relatively small number of classes which correspond to certain spatial geometric arrangments of the loops formed by the disulfide bonds and that each of these classes be represented by at least one conformation chosen at random, as explained in detail below.

As an illustration, consider a chain having only two long-range contacts, one (ae) connecting the two ends of the chain, and the other (bd) connecting two middle residues (Figure 1) (these contacts could be covalent or noncovalent interactions[39] but, for purposes of this discussion, we shall treat them as disulfide bonds). Two, and only two, loops are created by these contacts, viz., I and II.[40] There are 7 *simple* SGAL's (i.e., arrangements with no more than one penetration of a given loop) which are shown in Figure 1. Thus, in A of Figure 1, loops I and II are separated from each other. In B and B′ of Figure 1, loop I is threaded through loop II from two different sides of loop II. Arrangements C and C′ show two different ways of intertwining loops I and II. In D and D′ of Figure 1 loop II is threaded through loop I from two different sides of loop I. Even though A, B, B′, D and D′ represent five different SGAL's they are topologically equivalent because they can be interconverted without breaking any covalent bonds; C and C′, however, are not only different SGAL's from the other five, but also have a different topology because it would be necessary to break disulfide bond ae or bd, form this SGAL, and then remake these disulfide bonds to convert A, B, B′, D, or D′ to C or C′ (this is true provided that rotation around bond bd is not allowed). Each of the SGAL's in Figure 1 constitutes a different class of conformations. Since loop penetration is well-defined mathematically, the classes of SGAL's are well-defined for *any* arrangement of the two loops; it should be noted that there is a *range* of conformations within each SGAL class. It is important to point out that proteins are simpler systems than is implied by a mathematical representation of loops, as in Figure 1. First of all, it is unlikely that proteins will be threaded to any great degree.[39] Secondly, the stiffness of the backbone and the bulkiness of the side chains reduce the possible number of SGAL's below that for a corresponding mathematical curve. Thirdly, knotting of the polypeptide chain is unlikely to occur, and in fact has never been observed.[39,41]

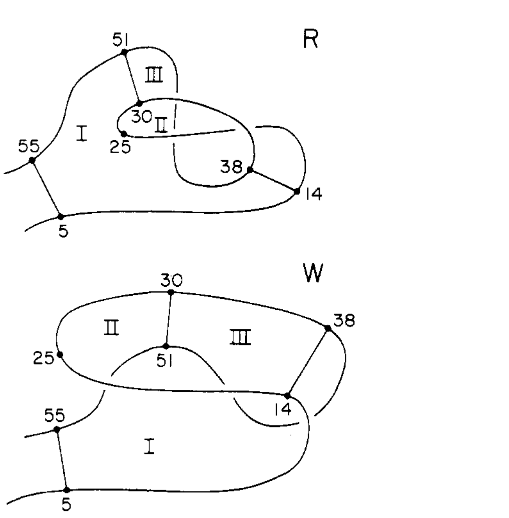

Figure 2. The three loops of BPTI. R, right SGAL, as in the native protein; W, wrong SGAL. Loop I is formed by disulfide bond 5–55, loop II by 14–38, and loop III by 30–51. In R, segment 14–30 (which is part of loop II) passes through loop III, whereas it does not in W.

In general, the number of classes depends on the length of the chain, the positions of the contacts, and the amino acid sequence (since the sequence determines the flexibility of the chain; also bulky side chains can prevent loop penetration). In the case of BPTI, there are three disulfide bonds, and hence three loops (Figure 2). It is more difficult to represent the various SGAL's of BPTI, as was done in Figure 1 for the hypothetical protein with two disulfide bonds. To the best of our knowledge, no theoretical treatment has been given of the number of possible classes of SGAL's for a given number of disulfide bonds. Therefore, we were able to define the classes of SGAL's for BPTI only by using the geometrical criteria underlying Figure 1 and by manipulating wire models (while paying attention to the SFM) to identify as well as possible the total number of SGAL's.

It should be noted that, unlike the loops of Figure 1, loops II and III (see Figure 2R which shows the correct SGAL) have a common segment, viz., 30–38. Loop I contains almost the whole molecule (5–55) but, for the sake of convenience, we prefer to treat the "pseudoloop" I′ (5, 14, 38, 30, 51, 55, 5), which is not a loop according to the conventional definition (see the discussion of loop II[40] in Figure 1). Loop I′ shares the same common segment, 30–38, with loops II and III. It should be pointed out that the remaining segments of these loops (viz., the segments that do not contain segment 30–38) are relatively small (comprising 16, 13, and 13 residues in loops II, III, and I′, respectively); therefore, these loops can penetrate each other (in the sense of Figures 1B, 1B′, 1D, and 1D′) only slightly. Hence, these penetrating structures are less likely to occur and, if they do, they do not differ much from similar nonpenetrating structures; therefore, they are not considered as separate SGAL's in the present case.

In order to find the number of SGAL's for BPTI, we first deduce the number of SGAL's for loops II and III. These loops can form five *unknotted* SGAL's: (a) the two loops are not threaded into each other, as in Figure 1A; (b,c) loop III is threaded through loop II from two different sides of loop II; (d,e) loop II is threaded through loop III from two different sides of loop III. For example, in the SGAL of the native structure, segment 14–30 of loop II is threaded through loop III, with segment 14–25 passing below segment 30–38, and segment 30–38 passing above segment 38–51, as shown in Figure 2R [this SGAL is denoted as d above]. For each of these five SGAL's, we found four of the following five different *unknotted* geometrical arrangments of loop I′ (a different four for each SGAL): (i) loop I′ is not threaded through loops II and III; (ii) loop I′ is threaded from one side of loop III; (iii) loop I′ is threaded from one side of loop II; (iv, v) loops II and III are threaded simultaneously by loop I′ from both sides. In SGAL d, however, loop II is twisted and, therefore, two of the above four arrangements are sterically not allowed because of the bulkiness of the side chains. Thus, we estimate that there are 18 simple SGAL's for BPTI. In Figure 2W, we show one of the wrong SGAL's; it is wrong in the sense that loops II and III do not have the same SGAL as the native structure.

Even if one member (conformation) is selected from each class of SGAL's, this is still a large number of starting conformations. This number cannot be reduced, in general, by observations[39,41] that loop penetration and threading do not occur frequently in proteins, because BPTI is an example (albeit an infrequently occurring one) of a threaded protein. The number of classes can, however, be decreased by imposing restrictions of the following types on the SFM. First of all, only compact members of the SGAL's should be considered so that the density of the model should resemble the experimental density. Secondly, since hydrophobic (hydrophilic) amino acids occur more frequently (infrequently) in the interior than on the surface of proteins, members of SGAL's with radial distributions of amino acid residues (around the center of mass) which deviate highly from the observed average distribution can also be excluded (for details, see ref 42–45).[46] The first criterion can be applied directly to the SFM, and the second criterion to the computer-generated version of the SFM. Imposing these restriction enables the effect of the solvent, which is difficult to treat rigorously, to be taken into account in an empirical manner.[46] In the present work, we did not impose any of these restrictions, nor did we use any statistical information from the known structures of proteins to restrict the dihedral angles of the amino acid residues to be located in helix, extended, or other states. The only restrictions on the dihedral angles were those imposed by steric hindrance and the SGAL's in the SFM. Such additional restrictions are being imposed in related work in progress.

It should be emphasized, however, that a large number of conformations belong to a given class of SGAL's. We select conformations at random in each class. Clearly, it is essential to have a good representation from the class of the native structure (which, of course, is unknown). In a first approximation, we may select one conformation from each class. In a second approximation, we may select more than one from each class. If two conformations from a given class differ significantly, it would be necessary to use the second approximation. The conformations selected in either of these approximations constitute starting conformations for energy minimization in the second stage.

In the second stage of our procedure the dihedral angles, corresponding to each of the starting conformations, are determined from the SFM and provided to the computer as the initial data for energy minimization. It should be pointed out that the structure generated by the computer generally differs from the one constructed with the help of the SFM for several reasons: (a) a reading error of ±5° in the values of the dihedral angles can change the structure of large chains drastically; (b) because of the weight of the SFM, the plastic connectors are distorted significantly; and (c) there are slight differences in bond lengths and bond angles between the CPK SFM and the model of amino acids used by ECEPP.[29] Because of this problem, an ORTEP stereoplot of each starting conformation is examined before minimization in order to make sure that the SGAL is not destroyed. If the SGAL is destroyed in the computer-generated conformation, we correct it as explained in the next section.

The energy function (ECEPP[29]) that is minimized consists of nonbonded and electrostatic interactions as well as torsional and artificial disulfide loop-closing potentials.[36,37] ECEPP constitutes a very detailed model of the protein which also takes all hydrogen atoms into account (the hydrogen bond potential, involving polar hydrogen atoms, is included in the nonbonded portion of the algorithm).

The energy is minimized with respect to two groups of variables[47] with the help of two minimization procedures. The first is MINOP,[48] which requires the calculation of gradients; with this procedure, the system is directed to the closest local energy minimum, whose energy can be very high. In such a case, we then apply a second procedure [systematic alteration of dihedral angles (SADA)] which searches for lower energy regions beyond that in which the system was trapped by MINOP. In SADA, each variable is changed sequentially, one at a time, in increments of *n* degrees within some predefined range around its current value. If a lower energy is obtained for a change in one of the dihedral angles, the altered value replaces the current one; if not, the current value is retained. The complete minimization procedure consists of several cycles of SADA, followed by MINOP, and the process is repeated until the reduction of the energy in one cycle of SADA or in 20 derivative calculations with MINOP is less than 0.05 kcal/mol.

It should be pointed out that the minimization procedure described above, which is based upon both MINOP and SADA, takes the molecule out of the region closest to the starting conformation and, therefore, the lowest attainable energy depends to some extent on the order in which MINOP and SADA are applied. Hence, in comparing the final energies obtained from two starting conformations, one must bear in mind that the results are not absolute in the sense that they depend on the minimization procedure used.

The calculations were carried out on a system consisting of an FPS AP-120B array processor and a Prime 350 minicomputer host.[49] Even though the AP is one of the fastest machines available at present, a complete energy minimization[47] with respect to all 267 dihedral angles of

Table I
Energy Values of the Six Structures (kcal/mol)[a]

conformation[b]	electrostatic	nonbonded	torsional	cystine torsional[c]	loop closing[d]	total
WI	54.8	2.25×10^8	108	687	2.85×10^4	2.25×10^8
WF	19.4	104	35.2	8.66	0.940	168
NI	−20.2	1.71×10^6	81.0	21.9	185	1.71×10^6
NF	−14.6	−221	45.8	8.53	2.03	−179
RI	15.3	2.20×10^{10}	148	3820	1.24×10^5	2.20×10^{10}
RF	12.1	18.9	49.8	11.7	1.80	94.3

[a] Definitions of the various interactions are given in ref 27. [b] See text for definitions of the six conformations. [c] This is the torsional energy about the C^β–S and S–S bonds. [d] This is the energy arising from the departure of the S–S bond distance and the nonbonded $C^\beta \cdots S$ distances from their standard values.

BPTI requires a considerable amount of computer time (see ref 49 for details). In this initial work to explore the procedure, we therefore minimized the energies of only three starting conformations: (a) a structure of BPTI that was partially refined by Swenson et al.[50] from the X-ray coordinates of Deisenhofer and Steigemann[51] (This initial "native" structure[50] is referred to as NI, and the final structure obtained from it by energy minimization is referred to as NF.); (b) a conformation obtained by use of the SFM, which had the same (i.e., right) SGAL as in NI, but different from NI in many respects (e.g., it did not have the C-terminal α-helix) (We refer to this conformation as RI, and to the final structure obtained from it by energy minimization as RF.); and (c) a conformation, also obtained by use of the SFM, but which had the wrong SGAL (This conformation is referred to as WI, and its energy-minimized structure as WF.). The right and wrong SGAL's of BPTI are illustrated in Figure 2.

The energy of NI is minimized primarily to see if the native structure is one of minimum energy in the ECEPP representation.[29] Also, the value of the energy of NF can then serve as a reference with which to compare the energies of RF and WF.[52] In addition, a comparison of the structures of RF and NF might provide information as to the validity of this approach, i.e., as to whether the imposition of the right SGAL constitutes enough restrictions so that energy minimization would bring it to the native structure. Comparison of the energies of RF and WF will indicate whether our assumption (that a starting conformation from the class of the native structure should lead to a lower energy than those obtained from other starting conformations) is valid.

Results and Discussion

Energy Minimization of NI. NI was obtained by Swenson et al.[50] by fitting a standard-geometry model to the X-ray coordinates[51] of BPTI and carrying out a limited energy minimization. Their energy function used united atoms for the hydrogens and carbons of nonpolar groups, rather than treating the hydrogens explicitly, as is done by ECEPP. Their limited energy minimization was carried out segment by segment, with only a small number of dihedral angles being allowed to vary in each segment, and the atoms were constrained to remain close to their experimental positions. Since we use ECEPP, minimize with respect to 155 or 196 of the maximum of 267 variable dihedral angles,[47] and do not impose constraints to keep the computed structure near the experimental one, it could not be determined prior to the minimization what extent of change would be imposed on NI by our procedure.

Pottle et al.[49] carried out the initial part of the minimization of NI, using MINOP.[48] They varied 154 dihedral angles at a time—all the backbone dihedral angles, ϕ_i and ψ_i (except ϕ of Pro, which is kept constant at −75° in ECEPP), and all of the side-chain dihedral angles, χ_i^1, except for alanine. The high energy of NI (~10^6 kcal/mol; see Table I) arises from contacts between atoms which exist in ECEPP but did not exist in the united atom model.[50] Their minimization required ~27 h of computer time, and their final energy was ~+100 kcal/mol. Most of this time was spent on the computation of the derivatives of the energy, ~3.5 min being required to compute 154 derivatives.[49]

Their final structure was used here as a starting one for further energy minimization. First, MINOP was used, and all 155 side-chain dihedral angles were varied;[47] this lowered the energy to −91 kcal/mol. Then 227 dihedral angles [all backbone dihedral angles (ϕ_i, ψ_i), except ϕ of Pro, and the side-chain dihedral angles χ_i^1, χ_i^2, and χ_i^3] were allowed to vary.[47] This decreased the energy further, to ~−110 kcal/mol. This was the lowest energy that could be obtained with MINOP *for this choice of variable dihedral angles*; *presumably*, the minimum of the local potential well has thus been reached.

At this stage, SADA was employed. Each dihedral angle was changed sequentially (in the order specified by the ECEPP algorithm) around its current value in increments of 1° within the range of ±5, then ±10, and then ±20°. If a lower energy was obtained for a change in one of the dihedral angles, the altered value replaced the current one. This procedure was applied three times to all variables sequentially, and the energy was lowered to −164 kcal/mol. Further use of MINOP lowered the energy to −179 kcal/mol. The total energy and its components are given in Table I. The electrostatic energy changed very little; in fact, it increased slightly between NI and NF. The torsional energies decreased slightly. A somewhat larger decrease was obtained for the (disulfide) loop-closing potential, which means that the geometry around the S–S bonds approached standard values more closely. The largest change in energy arose from the nonbonded (including hydrogen bond) interactions, due to the removal of atomic overlaps during minimization.

Some geometrical properties of NI and NF are given in Table II. R_g is the root-mean-square radius of gyration with respect to only the C^α atoms. RMS is the root-mean-square deviation of a given conformation from NI[52] with respect to only the C^α atoms:

$$\mathrm{RMS} = \left[\sum_{i=1}^{N} \frac{1}{N}(d_i - d_i')^2 \right]^{1/2} \qquad (1)$$

where d_i and d_i' are the distances between the ith pair of C^α atoms in NI and in the compared structure, respectively, and N is the total number of C^α pairs. Since only residues 1–55 were located in the experimental electron density map,[51] the root-mean-square deviation is computed only for residues 1–55. The contact maps, however, are given for 58 residues, with residues 56–58 added to complete the structure. $D_{5,55}$, $D_{14,38}$, and $D_{30,51}$ are the distances between the C^α atoms of the three disulfide-bonded pairs

Table II
Geometrical Parameters for the Six Structures (Å)[a]

conformation	R_g	RMS	$D_{5,55}$	$D_{14,38}$	$D_{30,51}$	$D(\text{S-S})_{5,55}$	$D(\text{S-S})_{14,38}$	$D(\text{S-S})_{30,51}$
WI	13.7	8.9	10.4	8.1	4.8	11.9	5.1	5.5
WF	13.2	8.3	5.4	6.0	6.2	2.0	2.0	2.0
NI	10.4	0	5.1	6.5	6.3	1.8	2.7	1.9
NF	10.6	0.5	5.8	5.9	6.6	2.0	2.0	2.1
RI	14.8	8.7	19.7	16.7	15.3	14.9	13.7	13.6
RF	11.7	5.9	6.3	5.7	6.9	2.1	2.0	2.0

[a] See text for definition of the parameters.

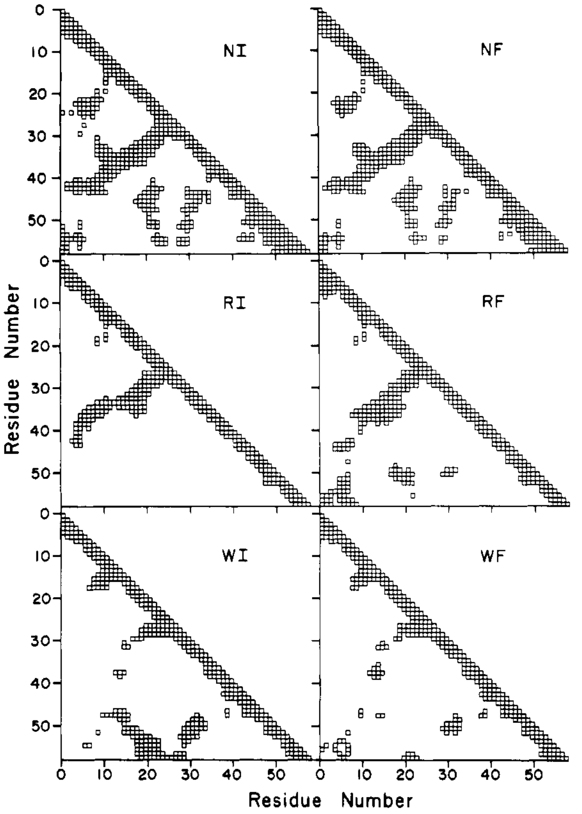

Figure 3. Contact map of the various conformations. The squares indicate pairs of C^α atoms whose distances of separation are ≤10 Å.

of residues. The corresponding distances, D(S–S), between the three pairs of S atoms are given in the last three columns of Table II. The parameters for NI and NF are close to each other. R_g is slightly larger in NF than in NI, which indicates that the optimized structure is more open. The largest deviation for the disulfide bonds is 0.7 Å (for $D_{5,55}$), which is close to the value of RMS = 0.5 Å; this means that the geometry around the disulfide bonds is about as good as that of the rest of the structure. The changes, $\Delta\phi$ and $\Delta\psi$, that occurred in the backbone dihedral angles are listed in Table III. Relatively large differences, $\Delta\phi$ and $\Delta\psi$, were obtained for the two ends of the chain, whereas the middle portion changed very little in minimization. The similarity between the two conformations is also demonstrated by comparison of their 10-Å contact maps (Figure 3) and their ORTEP stereodiagrams (Figure 4).

The facts that NI and NF are very similar (RMS = 0.5 Å for C^α atoms) and that NI is close to the X-ray structure (RMS = 0.6 Å for all backbone atoms; see Table I of ref 50) mean that NF is also close to the experimental conformation. This indicates that the native structure is one of minimum energy in ECEPP but the results of these computations do not enable us to decide whether this is the global or a local minimum.

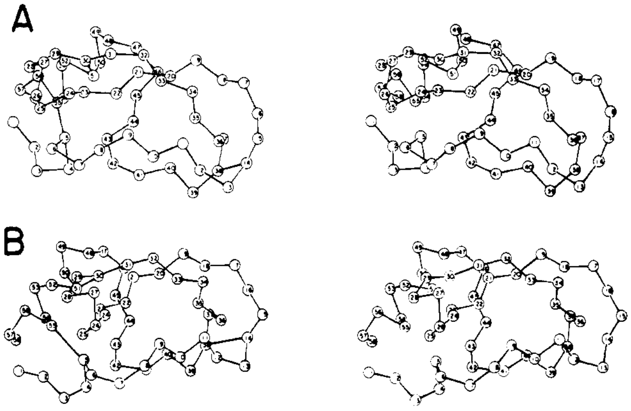

Figure 4. Stereoviews of virtural bond representations of (A) NI and (B) NF. The disulfide bonds connect half-cystine residues 5 and 55, 14 and 38, and 30 and 51.

Energy Minimization of SFM Structures. The SFM was fixed in the desired conformation by means of iron stands, rods, and wires. It takes several hours to read all the 267 backbone and side-chain dihedral angles from the dials of the model. In order to avoid errors, we have generally carried out this reading three times and then averaged the readings. As has been pointed out already, the structure generated by the computer differs from the one defined by the SFM mainly because of reading errors and the distortion of the SFM plastic connectors caused by the weight of the SFM. Therefore, an ORTEP stereodiagram of the computer structure was always examined to make sure that the SGAL defined by the SFM still existed in the computer structure. The energies of the two SFM structures illustrated in Figure 2, RI and WI, were minimized by the same procedure used for NI.

The SFM of RI was constructed to give the SGAL of the native structure, and the only other restrictions were the avoidance of short-, medium-, and long-range overlaps for both backbone and side-chain atoms. We did not assign any conformational state (helix, extended, etc.) to any residue nor did we try to pack the model at the right density. In fact, the dihedral angles obtained by this procedure differed substantially from the experimental X-ray ones. In this sense, our choice of RI is a random one in the class to which NI belongs.

The initial conformation of RI, generated by the computer from the set of dihedral angles provided by the SFM, did not have the correct SGAL because of the types of errors mentioned above; i.e., instead of segment 30–38 being above segment 14–25 (as illustrated in R in Figure 2), it was below. We directed segment 30–38 to lie above segment 14–25 by increasing ψ of Ala-27 by 25°. This change of ψ of Ala-27 not only produced the correct SGAL but also brought the disulfide-bonded pairs 5–55 and 14–38 (which previously were far apart) closer together. The resulting structure (RI) thus had the correct SGAL but still differed considerably from the original SFM structure,

Table III
Changes in the Backbone Dihedral Angles, $\Delta\phi$ and $\Delta\psi$ (in Deg), between the Initial and Final Energy-Minimized Structures

	NF - NI		RF - RI		WF - WI	
Residue	$\Delta\phi$	$\Delta\psi$	$\Delta\phi$	$\Delta\psi$	$\Delta\phi$	$\Delta\psi$
Arg 1	-40.021	-20.554	4.998	-18.064	-15.001	28.012
Pro 2	0.000	-19.003	0.000	9.088	0.000	52.038
Asp 3	-14.117	6.625	-11.927	1.882	-12.008	-16.017
Phe 4	-6.318	-0.727	56.922	-30.850	44.098	-31.066
Cys 5	-3.337	1.631	27.301	-21.792	-17.020	-8.238
Leu 6	-1.733	-0.454	22.037	1.333	11.881	-8.715
Glu 7	0.079	0.303	26.592	-11.437	-7.820	-1.882
Pro 8	0.000	-0.783	0.000	-11.979	0.000	2.133
Pro 9	0.000	0.723	0.000	-4.268	0.000	-1.255
Tyr 10	0.460	0.112	-0.615	-1.215	-1.401	2.356
Thr 11	-0.040	-1.421	8.995	-15.397	2.227	-7.589
Gly 12	0.186	1.210	3.023	-0.738	-1.687	-5.487
Pro 13	0.000	-0.452	0.000	1.066	0.000	7.916
Cys 14	-0.268	-1.355	-1.282	0.260	3.334	-9.728
Lys 15	-1.975	0.410	-2.119	2.979	-1.821	17.474
Ala 16	0.274	0.417	-1.549	-1.102	-3.727	3.091
Arg 17	0.342	0.284	11.014	-8.813	1.978	5.976
Ile 18	-0.132	0.348	4.112	-0.315	3.457	1.105
Ile 19	0.377	0.169	-4.050	-2.278	-4.200	0.940
Arg 20	-0.047	-0.413	0.548	2.272	0.871	0.674
Tyr 21	-0.688	0.416	1.678	-55.620	-0.032	1.882
Phe 22	-0.034	0.449	61.313	-1.031	0.587	5.131
Tyr 23	0.250	-1.888	0.763	-1.955	3.316	-0.238
Asn 24	-1.568	0.140	-1.240	-0.062	0.753	-2.693
Ala 25	-1.961	0.942	-2.267	2.190	2.369	-0.340
Lys 26	0.658	-1.886	3.831	5.622	0.113	-0.281
Ala 27	0.093	-0.540	1.310	-1.783	-0.511	0.362
Gly 28	-0.303	-1.161	0.301	-29.566	-0.643	1.535
Leu 29	-0.951	-0.737	0.944	3.437	-0.184	-0.766
Cys 30	-0.447	-0.475	2.105	-2.081	-0.395	0.652
Gln 31	-0.370	0.482	-1.534	1.108	-0.040	0.300
Thr 32	0.214	0.216	-6.522	-4.780	6.121	0.435
Phe 33	-0.002	1.214	9.922	7.946	2.601	-5.704
Val 34	0.658	0.351	1.370	-1.162	-0.673	-0.246
Tyr 35	0.507	-0.350	1.877	6.949	-0.255	0.445
Gly 36	-1.156	-0.509	2.132	-8.028	-0.492	-0.865
Gly 37	-0.979	0.004	5.251	2.834	0.147	1.653
Cys 38	-0.294	1.177	3.852	-4.789	0.892	0.069
Arg 39	1.336	1.536	-2.205	0.276	0.414	-1.664
Ala 40	0.688	-1.037	-3.296	-4.011	-0.730	0.063
Lys 41	-1.009	-0.031	-2.869	5.604	0.139	-1.136
Arg 42	0.710	0.763	4.594	-8.022	0.033	1.598
Asn 43	0.498	0.880	-0.970	0.714	0.739	2.936
Asn 44	0.941	0.294	3.091	4.127	1.340	0.412
Phe 45	0.248	0.771	3.087	8.458	0.265	0.985
Lys 46	0.658	0.584	9.287	-8.315	-0.042	-0.209
Ser 47	0.470	-0.119	-2.839	2.785	-0.083	6.779
Ala 48	-0.010	-0.316	3.729	2.625	0.741	-27.827
Glu 49	-0.765	0.795	1.514	-3.224	0.291	0.597
Asp 50	0.815	1.877	7.110	5.569	0.681	5.396
Cys 51	3.611	-0.525	-1.407	-4.327	-3.885	-6.549
Met 52	-1.295	1.455	0.773	-5.297	1.666	-4.929
Arg 53	-0.555	7.958	-5.779	2.959	-3.952	-20.574
Thr 54	1.898	3.493	-8.073	-40.251	-9.557	-13.108
Cys 55	6.695	21.865	-20.162	43.101	-20.457	-12.997
Gly 56	-9.420	-5.619	-20.860	-8.971	27.872	7.799
Gly 57	4.314	-1.444	5.020	65.024	-26.016	47.099
Ala 58	5.713	-18.995	-58.012	-75.015	16.003	-31.002

as can be seen from Table II and from the ORTEP stereoplot of Figure 5C [in the SFM, the S–S bond lengths are fixed by the CPK model at 2.08 Å, but the values of D(S–S) for the computer-generated version of RI were considerably greater than this distance; similarly for the distances between the C^{α} atoms of disulfide-bonded pairs]. The value of R_g was also very large. In other words, RI had a more open structure than it should have had.

The nonbonded energy of RI ($\sim 10^{10}$ kcal/mol) was much higher than the disulfide loop-closing energy ($\sim 10^5$ kcal/mol) and the other interaction energies. Thus, any minimization at this stage would be governed by the nonbonded rather than by the other interactions. Hence, minimization with respect to the backbone variables might have increased the distances between the disulfide-bonded pairs instead of bringing them together (thereby destroying the correct SGAL). Therefore, the energy was first decreased by varying *only* the side-chain dihedral angles; then the backbone variables were also allowed to change but without large movements of the backbone. In the initial stage of the calculations, SADA was applied to all of the 155 side-chain dihedral angles.[47] Each dihedral angle was allowed to change in increments of 1° in the range ±20° around its current value. In the next step, this increment and range were increased to 4° and ±180°, respectively. The energy decreased to $\sim 10^9$ kcal/mol. We then allowed all the backbone dihedral angles (ϕ_i, ψ_i) and side-chain dihedral angles χ^1 and χ^2 to vary (altogether 196 variables[47]), again using SADA. In order to avoid drastic geometrical changes of the backbone in the first few cycles, we confined the variation of these dihedral angles to a small range, ±4°, which is within the reading error (±5°). In later cycles, this range was increased to ±10°. When the energy was lowered to $\sim$5000 kcal/mol, we applied

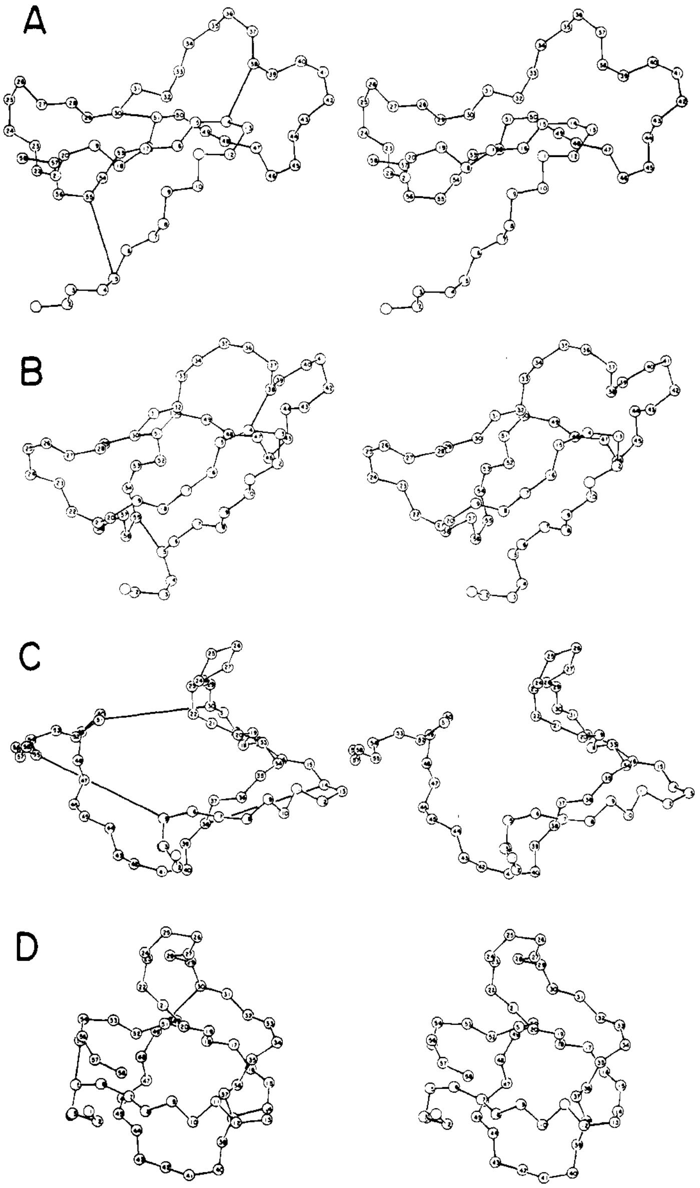

Figure 5. Stereoviews of virtual bond representations of (A) WI, (B) WF, (C) RI, and (D) RF.

MINOP, which was found to be very efficient in this range of energies; it lowered the energy to $\sim$1000 kcal/mol and we again applied SADA. It should be pointed out that we also tried to use MINOP when the energy was much higher, but the system was always trapped in a local minimum in contrast to the computation with NI, where MINOP appeared to be more efficient; we therefore had to apply SADA in the high-energy region. The lowest energy that could be achieved by the two minimizers was 94 kcal/mol, which is a relatively high energy.

The computer time required for this minimization was approximately 30 h. Table III reveals that the largest changes in backbone dihedral angles occurred for the residues of the two ends of the chain (as was found for NI). The changes in the dihedral angles in the *middle* of the chain, however, were larger than those obtained for NI.

Table II reveals that R_g of the optimized structure RF is $\sim$3 Å less than that of RI, but it is still higher than the value of R_g of NI, which indicates that the density of RF is smaller than that of NI. As seen in Table II, the disulfide bond distances are close to those of NI (the differences between the D's are much smaller than the value of RMS—5.9 Å); these differences do not exceed 1 Å, and in most cases are much less. A comparison of the ORTEP stereodiagrams of NI (Figure 4A) and RF (Figure 5D) reveals that the SGAL is the same in the two structures. Two of the turns are almost in the same places for the two structures, whereas the third turn occurs in residues 36–40 in NI but is shifted to residues 39–44 in RF. The α-helix which exists in NI near the C terminus, and which was not imposed on RI, also does not appear in RF. Other differences, in the short- and medium-range structure, exist between the two conformations; these can be seen from the two ORTEP stereodiagrams. Another way of comparing these structures is by means of the contact maps (Figure 3).

We also minimized the energy of another structure, WI, which did not have the same SGAL as in NI. This SGAL is represented in Figure 2B. The energy of WI was minimized by using the same procedure that was applied to RI. The minimizers succeeded in bringing the S$\cdots$S pairs close together as in NI and also in decreasing R_g slightly. Conformation WF, however, is much less dense than NI and less similar to NI than is RF. This is demonstrated by comparisons of the ORTEP stereodiagrams (Figure 5B, 4A, and 5D) and of the contact maps (Figure 3). Here also the minimization procedure affected primarily the dihedral angles of residues located at the two ends of the chain. The energy of WF is larger than that of RF by $\sim$75 kcal/mol, in accordance with our assumption (see Introduction) that a conformation representing the class of the native structure should lead (by energy minimization) to a structure with lower energy than structures belonging to the other classes. One must bear in mind, however, that even for WI and RI, the results of the minimization are not absolute in the sense that they depend on the specific minimization procedure used. It should also be pointed out that the present test of our assumption is based on the minimization of the energy of only two conformations (in which we find the energies to be in the order NF < RF < WF); additional computations of this kind will be required to establish the validity of this assumption.

Conclusions

Several conclusions can be drawn from this work.

(1) The SFM was found to be useful in defining conformations with certain long-range structural properties (i.e., S–S bonds and different SGAL's). The SFM, however, is not accurate enough and this inaccuracy generally leads to deviations of the structure generated by the computer from the SFM structure. In fact, some of the structural changes imposed by the energy minimization process were "corrections" of the deviations mentioned above. For example, the S–S bonded half-cystine pairs that were always found to be located too far from each other in the computer structures were brought back by the loop-closing potential in the energy minimization procedure to the correct distances, as in the SFM structure. It therefore seems to be very important to use, in addition to the SFM, a computer graphics system which would be helpful in defining better starting conformations and in facilitating control over the minimization process.

(2) Energy minimization of NI, *without* the use of any constraints to keep it close to the observed structure, led to a conformation that is structurally almost unchanged. This means that the native structure of BPTI corresponds to a minimum (with a negative value) in the energy function ECEPP. It is, however, not clear whether this minimum of ECEPP is also the global minimum.

(3) SADA and MINOP were not able to cause any drastic structural change in the compact starting conformations. The minimizers also failed to pack the molecule tightly

enough, and did not lead to the C-terminal α-helix, when this information was not included in the starting conformation. All of this means that energy minimization is very limited in bringing about significant changes in compact structures; therefore, most of the structural features of the native molecule should be included in the starting conformation. An SFM conformation which has the correct SGAL (such as RI) led to a structure resembling NI in overall shape, but not in the detailed short-range structure. To achieve agreement with NI in both the short- and long-range structures, more restrictions on the starting conformations, such as the correct experimental density, will have to be added. Other structural information derived from the known structures of proteins may also be used. In order to apply such restrictions efficiently, a computer graphics system would seem to be a very useful tool.

(4) We assumed that the starting conformation representing the class of the native structure would lead to lower energy than other starting conformations. The results of our minimizations confirm this assumption. One has to bear in mind, however, that energy minimization beyond the closest minimum depends on the pathway of minimization and, in this sense, the results are not absolute. Also, it will be necessary to minimize the energies of additional structures in order to test this assumption more completely.

(5) SADA was found to be efficient in lowering the energy in all ranges of energy. MINOP, on the other hand, was inefficient at high energies, since the system was trapped in high-energy local minima. It seems to be important to check this problem carefully and to test other minimizers as well. It should be pointed out that the computer programs of both minimizers (except for the computation of gradients) have not yet been adapted for use on the AP. The speed of computation can still be increased by a factor of 4–5 by optimizing the code for the gradient routine[49] and transferring the rest of the minimizing procedures to the AP.

(6) The definition of the SGAL's introduced in this initial study is intuitive. Work is currently in progress to try to develop a computer algorithm for the definition of SGAL's which will be useful for the estimation of the total number of classes. We are also considering the introduction of additional constraints that are being imposed on the SFM structure (which has the correct SGAL) in order to fold the protein successfully by energy minimization.

Acknowledgment. We are indebted to Dr. J. A. Yankeelov, Jr., for providing the tools and dials that were used with the CPK space-filling models. We also thank M. S. Pottle, S. Rumsey, and R. W. Tuttle for help with the computations and G. Némethy, M. Oka, M. S. Pottle, and H. Wako for helpful comments on this manuscript.

Note Added in Proof. These calculations have been extended by incorporating short-range restrictions in the present protein-folding algorithm. The root-mean-square deviation was thereby reduced to 4.4 Å [Meirovitch, H.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.*, in press].

Supplementary Material Available: Cartesian coordinates and dihedral angles of NI, NF, RI, RF, WI, and WF (114 pages). Ordering information is given on any current masthead page.

References and Notes

(1) This work was supported by research grants from the National Science Foundation (PCM79-20279, PCM77-09104, and PCM79-18336) and from the National Institute of General Medical Sciences of the National Institutes of Health, U.S. Public Health Service (GM-14312 and GM-25138).

(2) (a) Present address: Chemical Physics Department, Weizmann Institute of Science, Rehovoth, Israel. (b) To whom requests for reprints should be addressed.

(3) Sela, M.; White, F. H., Jr.; Anfinsen, C. B. *Science* **1957**, *125*, 691.

(4) White, F. H., Jr.; Anfinsen, C. B. *Ann. N.Y. Acad. Sci.* **1959**, *81*, 515.

(5) White, F. H., Jr. *J. Biol. Chem.* **1960**, *235*, 383.

(6) Anfinsen, C. B.; Haber, E.; Sela, M.; White, F. H., Jr. *Proc. Natl. Acad. Sci. U.S.A.* **1961**, *47*, 1309.

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

(8) Gō, N.; Scheraga, H. A. *Macromolecules* **1978**, *11*, 552.

(9) McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature* (*London*) **1977**, *267*, 585.

(10) The multiple-minimum problem *has* been solved for small open-chain and cyclic peptides, and for synthetic-peptide models of fibrous proteins.[11] These procedures, however, are not applicable to globular proteins, and other methods (as, e.g., the one to be discussed in the text) must be used instead.

(11) Scheraga, H. A. "Proceedings of the Fifth American Peptide Symposium"; Goodman, M., Meienhofer, J., Eds.; Wiley: New York, 1977; pp 246–56.

(12) Unlike energy minimization, Monte Carlo is not a deterministic procedure; therefore, there is always a finite probability to escape from a potential well. Such escape from local potential wells has indeed been demonstrated for small oligopeptides.[8] For large compact structures, however, this probability is very small.

(13) Scheraga, H. A. *Chem. Rev.* **1971**, *71*, 195.

(14) Némethy, G.; Scheraga, H. A. *Q. Rev. Biophys.* **1977**, *10*, 239.

(15) An approach of intermediate simplicity between that using a complete model of a protein and the highly simplified models discussed here was taken by Gibson and Scheraga. Gibson, K. D.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1967**, *58*, 420. *Ibid.* **1967**, *58*, 1317. *Ibid.* **1969**, *63*, 9, 242.

(16) Levitt, M.; Warshel, A. *Nature* (*London*) **1975**, *253*, 694.

(17) Levitt, M. *J. Mol. Biol.* **1976**, *104*, 59.

(18) Kuntz, I. D.; Crippen, G. M.; Kollman, P. A.; Kimelman, D. *J. Mol. Biol.* **1976**, *106*, 983.

(19) Burgess, A. W.; Ponnuswamy, P. K.; Scheraga, H. A. *Isr. J. Chem.* **1974**, *12*, 239.

(20) Chou, P. Y.; Fasman, G. D. *Biochemistry* **1974**, *13*, 211, 222.

(21) Fasman, G. D.; Chou, P. Y.; Adler, A. J. *Biophys. J.* **1976**, *16*, 1201.

(22) Maxfield, F. R.; Scheraga, H. A. *Biochemistry* **1976**, *15*, 5138. *Ibid.* **1979**, *18*, 697.

(23) Tanaka, S.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 142, 159, 168, 812. *Ibid.* **1977**, *10*, 9, 305.

(24) This problem, inherent in short-range models, has been overcome to some extent by Dunfield and Scheraga,[25] who used a nearest-neighbor Ising model (based on empirical potential energies rather than on X-ray data) and computed *individual* values of (ϕ, ψ) rather than *ranges* of such values.

(25) Dunfield, L. G.; Scheraga, H. A. *Macromolecules* **1980**, *13*, 1415.

(26) Tanaka, S.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1975**, *72*, 3082. *Ibid.* **1977**, *74*, 1320.

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

(28) Wako, H.; Scheraga, H. A. *Macromolecules* **1981**, *14*, 961.

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

(30) Crippen, G. M. *J. Comput. Phys.* **1975**, *18*, 224.

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

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

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

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

(35) Yankeelov, J. A., Jr.; Coggins, J. R. *Biopolymers* **1972**, *11*, 707.

(36) While ECEPP uses an artificial potential to form disulfide-bonded loops, it is of interest that the loop-closure problem has also been solved by a procedure which guarantees exact loop closure.[37] This is, however, a time-consuming calculation which is practical only for small loops such as those in cyclic oligopeptides.

(37) Gō, N.; Scheraga, H. A. *Macromolecules* **1970**, *3*, 178.

(38) Némethy, G.; Scheraga, H. A. *Biopolymers* **1965**, *3*, 155.

(39) Connolly, M. L.; Kuntz, I. D.; Crippen, G. M. *Biopolymers* **1980**, *19*, 1167.

(40) Strictly speaking, by the commonly used definition, II (formed by abdea) is *not* a loop because ab and de are two separated noncontinuous segments of the chain, i.e., the second loop is I + II (formed by abcdea); however, it is convenient to describe

the SGAL's in terms of the relative positions of I and II, rather than of I and I + II and, for this reason, we refer to II as a loop in the following discussion.

(41) Klapper, M. H.; Klapper, I. Z. *Biophys. J.* **1980**, *32*, 216. *Biochim. Biophys. Acta* **1980**, *626*, 97.

(42) Krigbaum, W. R.; Komoriya, A. *Biochim. Biophys. Acta* **1979**, *576*, 204.

(43) Meirovitch, H.; Rackovsky, S.; Scheraga, H. A. *Macromolecules* **1980**, *13*, 1398.

(44) Meirovitch, H.; Scheraga, H. A. *Macromolecules* **1980**, *13*, 1406.

(45) Meirovitch, H.; Scheraga, H. A. *Macromolecules* **1981**, *14*, 340.

(46) BPTI and HIPIP are two proteins with radial distributions of hydrophobic and hydrophilic amino acids that differ from the general behavior.[43,44] Therefore, the average distribution cannot be used in these cases.

(47) Minimization of the energy with respect to all 267 backbone and side-chain dihedral angles is very time-consuming. We found it more efficient to minimize in two cycles. In the first, the energy was minimized with respect to all backbone dihedral angles and all side-chain dihedral angles χ^1 and χ^2 (which have the largest effect on the orientation of the side chain), i.e., 196 variables. In the second cycle, the energy was minimized with respect to all 155 side-chain dihedral angles. In one trial, 227 dihedral angles were varied (ϕ, ψ, χ^1, χ^2, and χ^3), but in all other trials only 155 or 196 dihedral angles were varied.

(48) Dennis, J. E.; Mei, H. H. W. Technical Report No. 75-246, 1975, Department of Computer Science, Cornell University, Ithaca, N.Y.

(49) Pottle, C.; Pottle, M. S.; Tuttle, R. W.; Kinch, R. J.; Scheraga, H. A. *J. Comput. Chem.* **1980**, *1*, 46.

(50) Swenson, M. K.; Burgess, A. W.; Scheraga, H. A. In "Frontiers in Physicochemical Biology"; Pullman, B., Ed.; Academic Press: New York, 1978; p 115.

(51) Deisenhofer, J.; Steigemann, W. *Acta Crystallogr, Sect. B* **1975**, *31*, 238.

(52) We compare our calculated structures with NI rather than with NF because NI is closer than NF to the experimental X-ray structure. Nevertheless, the energy of NF serves as a reference for the lowest energy attainable by our procedure.

Differential Geometry and Polymer Conformation. 3. Single-Site and Nearest-Neighbor Distributions, and Nucleation of Protein Folding[1]

S. Rackovsky[2a] and H. A. Scheraga*[2b]

Biophysics Department, Weizmann Institute of Science, Rehovot, Israel, and Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853.

Received February 17, 1981

ABSTRACT: A differential geometric representation of polymer chains is used to study protein backbone structure on the four- and five-C$^\alpha$ length scales. The analysis of the distribution of four-C$^\alpha$ units in the curvature–torsion (κ, τ) plane reveals that there are islands of structure which occur with greater-than-random probability. These correspond to the well-known types of backbone structure—extended (E), right-handed α-helical (A_R), and flat-bend (A_0) structure. It is shown that there are three distinct types of extended four-C$^\alpha$ structure—left-handed twisted (E_L), right-handed twisted (E_R), and nearly flat (E_0). The E_0 and E_L structures form a structural continuum, but the E_R region is separated from this continuum by a region of low occupation. The A_R and A_0 regions also form a continuum. These high-frequency islands are distributed throughout the occupied region of the (κ, τ) plane and include structures of all types except that in the A_L (left-handed α-helical) region. The high frequency of occurrence of these structures suggests that they are of low energy and therefore likely to occur in nucleation structures in the folding of the denatured molecule. It follows from the fact that these high-frequency structures occur throughout the occupied region, however, that there is *relatively low selectivity in nucleation on the four-C$^\alpha$ scale*, with structures representative of the entire occupied region being potential nuclei on the four-C$^\alpha$ scale. Extension of the analysis to the five-C$^\alpha$ scale shows that the high-frequency structures on this scale are made up of combinations of the high-frequency four-C$^\alpha$ structures and that *a higher degree of selectivity in nucleation appears on the five-C$^\alpha$ length scale* than on the four-C$^\alpha$ scale. Substantial differences are noted in the frequency of occurrence of the various combinations. Within the extended region, for example, the most frequently occurring structures are E_0E_0, with E_RE_R being the least frequent. A study of the correlation between nearest-neighbor four-C$^\alpha$ structures reveals that certain *pairs* of four-C$^\alpha$ structures have a low probability of occurring. A_0A_R and A_RA_0 have a low positive correlation, while E_XA_R and A_RE_X (where X is R, L, or 0) have a negative correlation, indicating that these structures tend to *avoid* pairing. Five-C$^\alpha$ components of nucleating structures are likely to be those which both occur with high frequency and show positive correlation between their four-C$^\alpha$ components. Of these, the most frequently occurring are E_XE_X and A_RA_R, which are repeating (regular) structures. The A_0 four-C$^\alpha$ structure plays an important role in nucleation, despite its relative numerical infrequency, because it is the principal four-C$^\alpha$ structure which forms potential five-C$^\alpha$ nucleating structures which are nonregular. Larger nuclei, which are not considered explicitly in this paper, presumably arise as additional four-C$^\alpha$ units associated with five-C$^\alpha$ nuclei already present in the folding chain.

I. Introduction

One of the significant aspects of protein architecture is the presence, in molecules with widely differing function and amino acid sequence, of certain well-defined structural features. Historically, the first such features to be noted were the α helix and the extended strand (which associates with other strands to form the β sheet). These two structures are particularly easy to observe and classify because they are built up of repeating units (residues) with similar conformation. They are therefore capable of occurring, and of being observed, on a wide variety of backbone length scales, from a single residue to tens of residues.

It later became clear that there is at least one important structural feature which occurs on a single, well-defined length scale. This is the bend, or chain reversal, whose