ELSEVIER

Biophysical Chemistry 59 (1996) 329–339

Biophysical Chemistry

Recent developments in the theory of protein folding: searching for the global energy minimum

Harold A. Scheraga

Baker Laboratory of Chemistry, Cornell University, Ithaca, NY 14853-1301, USA

Abstract

Statistical mechanical theories and computer simulation are being used to gain an understanding of the fundamental features of protein folding. A major obstacle in the computation of protein structures is the multiple-minima problem arising from the existence of many local minima in the multidimensional energy landscape of the protein. This problem has been surmounted for small open-chain and cyclic peptides, and for regular-repeating sequences of models of fibrous proteins. Progress is being made in resolving this problem for globular proteins.

Keywords: Proteins; Folding; Global energy minimum; Statistical mechanics; Computer simulation; Multiple minima

1. Introduction

After Anfinsen [1] demonstrated the spontaneity of protein folding, introducing the hypothesis that native proteins (that are not post-translationally processed) are in their thermodynamically most-stable state, we began to try to compute the structures of proteins [2] from a knowledge of their amino acid sequences. Around the same time, Ramachandran et al. [3] and Schellman and Schellman [4] examined the stereochemical constraints on the conformations of terminally blocked alanine, and we explored the statistical mechanical foundations of the interactions in a polypeptide chain [5–10].

More recently, use has been made of lattice models, analytical theories, and computer simulation to investigate some fundamental aspects of the pathways of protein folding [11–39]. For example, the influence of amino acid sequence and the nature of the potential energy functions on the type of folding transition have been elucidated [27,35].

To compute the native structure of a protein, it is necessary to (i) generate an arbitrary starting conformation, (ii) compute its conformational energy (including entropy and hydration contributions), and (iii) locate the global minimum of the conformational energy. Adequate procedures are available for steps (i) and (ii) and for minimizing the conformational energy [40]. The minimization procedure, however, leads only to the local minimum closest to the starting conformation, rather than to the global minimum; this is the multiple-minima problem [40–42]. It is, therefore, necessary to have efficient methods for searching conformational space to locate the lowest minimum among all those in the whole space. In this paper, dedicated to Bill Harrington, we discuss some of the methods that we have developed to facilitate such a search of conformational space.

SSDI 0301-4622(95)00126-3

2. Summary of methods for searching conformational space

The following methods have been developed in our laboratory for searching conformational space.

1. Build-up procedure [43–45]
2. Self-consistent electrostatic field (SCEF) method [46]
3. Monte-Carlo-plus-Minimization (MCM) method [47,48]
4. Electrostatically driven Monte Carlo (EDMC) procedure [49–53]
5. Adaptive importance sampling Monte Carlo procedure [54–56]
6. Relaxation of dimensionality [57,58]
7. Pattern-recognition importance-sampling minimization (PRISM) [59]
8. Diffusion equation method (DEM) [37,38,60–63]
9. Self-consistent multi-torsional field (SCMTF) method [64–66]
10. Minimization-in-a-confined-space approach (CSA) [67–69]
11. Lattice neural network minimization (LNNM) [70].

These methods surmount the multiple-minima problem for small open-chain and cyclic peptides, and for regular-repeating sequences of fibrous proteins analogous to collagen and silk [42]. For applications to globular proteins, the DEM seems to be the most promising. Examples of applications of these methods are given below.

Other methods, such as simulated annealing [71,72], have been proposed for searching conformational space. However, while this method has been shown to be a useful procedure in constrained problems [73], it has been found to be inapplicable in unconstrained Monte Carlo calculations, using the pentapeptide Met-enkephalin as an example [74].

3. Build-up method

In this method, one starts with the low-energy structures of single residues, and uses these to build up low-energy structures of dipeptides, tripeptides, etc., carrying out energy minimization at each stage. This requires storage of *many* (backbone and side-chain) low-energy conformations. The conformational states that are stored are ordered according to their energies, taking hydration into account with a solvent-shell model. As the peptides become longer and longer, long-range interactions alter the (energetic) order of their conformations. The variety of structures being stored at each stage have conformations that allow such long-range interactions to come into play, as residues are added to the growing chain. From a practical point of view, this procedure and others have been applied to polypeptides as large as 25–30 residues. These 25–30 residue segments are then "stitched" together to build up even larger structures.

This method has been applied to several linear and cyclic polypeptides. In the case of the linear pentapeptide, [Met]-enkephalin, the same lowest-energy conformation [47,48] has been obtained by the build-up and by other methods described below; this attests to the validity of the several methods employed. However, other higher-energy structures pack better in crystals because of the presence of intermolecular hydrogen bonds [75]. Consequently, as shown by computations on several crystalline forms of enkephalin [75], the observed crystal structures are favored over hypothetical-packed structures formed from the global minimum structure of the isolated molecule. On the other hand, the conformations in the crystal have higher energies (as isolated molecules, where they are deprived of their intermolecular interactions) than that of the global-minimum structure of the isolated molecule.

If a peptide is cyclic, there is an additional constraint, namely that the ring must close exactly. In addition, cyclic peptides are probably less susceptible to crystal packing interactions than are flexible linear ones. The computed [76,77] structure of the cyclic decapeptide, gramicidin S, is in agreement with the subsequently obtained X-ray [78] and NMR [79,80] structures.

The build-up method has also been applied to fibrous and globular proteins. Collagen is an example of a fibrous protein which involves interchain association to form a triple-stranded coiled-coil structure. Conformational energy calculations have been carried out on several synthetic poly(tripeptide) analogs, poly(Gly-X-Y), of collagen [81,82]. Because of the regularity conditions imposed on each tripeptide in the computations, the number of degrees

of freedom was small, so that, again, the multiple-minima problem was surmounted by an adequate coverage of conformational space, using the build-up procedure. The computations indicated that poly(Gly-Pro-Pro), poly(Gly-Pro-Hyp), and poly(Gly-Pro-Ala) form stable, triple-stranded, coiled-coil, collagen-like structures, whereas poly(Gly-Ala-Pro) does not, all in agreement with the experiment.

After completion of the calculations [81] on poly(Gly-Pro-Pro), it was learned that Okuyama et al. [83] had carried out a single-crystal X-ray structure analysis of $(Pro\text{-}Pro\text{-}Gly)_{10}$. Our calculated structure is in agreement with theirs, with a root mean square (r.m.s.) deviation of 0.3 Å for all (non-hydrogen) atoms, based on a comparison between the X-ray coordinates (kindly provided to us by Professor M. Kakudo) and our computed ones.

Similar calculations have been carried out for the stacked β-sheet structure of poly(Gly-Ala) [84], a model for silk fibrin.

The build-up procedure has also been applied to the globular protein, bovine pancreatic trypsin inhibitor [85]. In combination with a limited number of simulated (NMR) distance constraints (3.3 restraints per residue), the structure of this protein has been computed; the backbone atoms of the resulting structure have an r.m.s. deviation of 1.19 Å from an idealized model of the X-ray structure.

4. Self-consistent electrostatic field (SCEF) method

In this SCEF procedure, we make an initial approximation by neglecting all components of the total energy except the electrostatic, and assume that each residue must have optimal electrostatic energy, i.e., the dipole moment of each residue must be optimally aligned in the electrostatic field created by the whole molecule. If it is not, we change the orientation of the dipole moment (of each residue, successively) to improve its electrostatic energy. Since this involves a local movement (in the field of the whole molecule), it is computationally very fast. Then the energy of the whole molecule (taking all interactions, not only electrostatic, into account) is minimized, and the whole procedure is repeated iteratively.

Thus far, we have tested this procedure on a 19-residue poly(L-alanine) chain with acetyl- and *N*-methyl amide terminal blocking groups. The global minimum of this structure is presumably a right-handed α-helix. We started with conformations very far from the helical conformation, and in trivially short computation time achieved the global minimum [46]. Unlike the usual minimization procedures, which make small changes in the dihedral angles, this procedure can make very large changes (even 100°–200°) in these independent variables.

5. Monte-Carlo-plus-Minimization (MCM) method

To overcome the inefficiency of Metropolis Monte Carlo, which searches all of the conformational space very slowly, we have devised a procedure to move rapidly through the space of local minima [47,48]. The energy of a random-starting conformation is minimized. Then a random change (selected in the range $0–2\pi$) is made in several randomly chosen dihedral angles, and the energy of this new conformation is minimized. The Metropolis criterion is used to decide whether to accept this new minimum, and the procedure is then iterated. In runs with 18 random starting conformations of the pentapeptide enkephalin, they all converged to the same global minimum [47,48] that had been obtained with the build-up procedure.

6. Electrostatically driven Monte Carlo (EDMC) procedure

The EDMC procedure incorporates the best features of the SCEF and MCM methods, combined with random conformational changes to simulate the effect of thermal motion [49]. This technique analyzes a given conformation (the current one), producing an electrostatic diagnosis based on the orientations of the dipole moments of the protein with respect to the local electric field. This diagnosis is used in combination with a random sampling technique to generate new conformations each of which is subjected to conventional energy minimization to reach a local energy minimum. This local minimum

is compared with that corresponding to the current conformation with the aid of the Metropolis criterion. Each time that a conformation is accepted, it replaces the current one and is subjected to an identical analysis. If all the electrostatic diagnoses fail to produce an acceptable conformation, and this situation remains unalterable after generating a significant number of random conformations, the process is forced to choose one of the conformations generated previously (but rejected) and to accept that one as the subsequent current conformation. This procedure is equivalent to a perturbation due to thermal effects. The method has been applied to poly(L-alanine) [49], Met-enkephalin [50], the 20-residue membrane-bound portion of melittin [51], decaglycine [52], and bovine pancreatic trypsin inhibitor [53].

7. Adaptive importance sampling Monte Carlo procedure

Another procedure to overcome the inefficiency of Metropolis Monte Carlo is adaptive importance sampling [54–56]. In this technique, the partition function (and quantities derived from it, such as the probability of a given conformation) is evaluated by continually upgrading the distribution function (ultimately, to the Boltzmann distribution) to concentrate the sampling in the region(s) where the probabilities are highest. Tests with enkephalin led to a low-energy structure.

8. Relaxation of dimensionality

Vanderbilt and Louie [72] developed an annealing approach in which the temperature of the system is raised (when the minimization becomes trapped in a local minimum) and a Monte Carlo procedure is carried out to allow the system to escape from the local potential well. We have developed a method for relaxing a system, not by raising the temperature but by raising the dimensionality of the space [57,58]. In a higher dimensional space, there are many more degrees of freedom in which the atoms can move about, making it easier to adjust to a low-energy conformation. Many potential barriers in three dimensions do not exist in higher dimensions. Our method starts from a very low-energy high-dimensional conformation and obtains a low-energy three-dimensional structure from it by gradual contraction of the dimensionality. The contraction in dimensionality is achieved by use of Cayley–Menger determinants, of which we have derived a simplified form [57,58,86].

In using this procedure, the energy-minimization problem is recast in terms of distances as the primary variables. In this respect, there is a similarity to the distance-geometry approach of Crippen [87] and Braun et al. [88]; however, our method differs from these in the manner in which the distances are used. Each distance variable is initially set to its own minimum-energy value subject to whatever geometric constraints may be imposed on it. This starting set of distances is then at a global-energy minimum if the dimensionality is not a consideration, i.e., this is a lower bound on the energy. But such a structure is not embeddable in a three-dimensional space. To obtain a realizable three-dimensional structure, a penalty function is added to the objective function to be minimized. This penalty function consists of the Cayley–Menger five- and six-point determinants. The purpose of these is to force the 4- and 5-D volumes of the structure to zero (to satisfy the necessary and sufficient condition that a structure be embeddable in a three-dimensional space). A penalty function consisting of upper and lower bounds on the distance is also added. These bounds are obtained from covalent geometric constraints but may optionally contain bounds from other theoretical or experimental considerations. By steadily increasing the weight of the determinants, the distances are forced to three-dimensionality, and the three-dimensional global energy minimum is approached from below rather than from above. The expectation is that, as the dimensionality is gradually reduced, the structure, having started from a high-dimensional global minimum, will evolve into the three-dimensional global energy minimum. Preliminary results [57] for a virtual-bond pentapeptide and for full-atom representations of several terminally blocked amino acids are encouraging, and application of the method to enkephalin [58] led to the same global minimum obtained by the other methods described above.

The fact that the several methods discussed here

all lead to the same global-minimum structure and energy of enkephalin, when the same potential function (ECEPP, Empirical Conformational Energy Program for Peptides [89–92]) was used, attests to the efficacy of each of these procedures.

9. Pattern-recognition importance-sampling minimization (PRISM)

Pattern recognition techniques are used to predict a series of probable backbone structures, whose energies are then minimized to locate the global minimum [59]. This is essentially a build-up procedure with probabilities instead of energies, saving the more-expensive energy minimization for the last stage of the procedure. The (ϕ, ψ) map of each residue is divided into four regions (α, ε, α^* and ε^*), and all possible tripeptides from a properly selected set of X-ray structures from the Brookhaven Protein Data Bank [93] are collected and grouped according to conformation (e.g., $\alpha\alpha\alpha$, $\alpha\varepsilon\varepsilon$, $\alpha\varepsilon\varepsilon^*$, etc.). The pattern recognition procedure uses amino acid properties [94] to map peptide sequences into a multivariate property space. Particular tripeptide conformations tend to map to particular regions of the property space. These regions are represented by multivariate Gaussian distributions, where the parameters of the distributions are determined from tripeptides in the Protein Data Bank. These data are then used to calculate the probability that each tripeptide in a protein under study has a given conformation.

The polypeptide chain is built up from the N-terminus, fitting the most probable tripeptide conformations together, one tripeptide at a time, allowing for proper overlap of the tripeptides. As the build-up proceeds, the probabilities of the growing chain (conformation) are calculated, and only the 1000 most probable are retained. Thus, when the C-terminus is reached, there are 1000 different predictions of the backbone structure of the protein, sorted in order of decreasing probability.

The symbolic representation (in terms of the regions, α, ε, α^*, ε^*) of the conformation of a protein is converted to a dihedral angle representation by randomly generating values of ϕ and ψ in each of the assigned regions from appropriate probability distributions. A bivariate (2-D) Gaussian distribution parameterized on values of (ϕ, ψ) from the known X-ray structures [93] is used, together with standard techniques for generating random numbers from Gaussian distributions. Several such random structures are generated for each backbone prediction, and the energy of each of them is minimized. The lowest-energy structure is taken to represent the backbone prediction. The aforementioned probabilities serve to reduce, to a manageable size, the set of conformations whose energies have to be minimized. This procedure has been tested on the 36-residue avian pancreatic polypeptide, and the computed lowest-energy structure [59] agrees reasonably well with the X-ray structure [95,96].

10. Diffusion equation method (DEM)

We have recently developed another algorithm to search for the global minimum of a potential energy function in the conformational analysis of molecules [37,38,60–63]. The algorithm is based on the deformation of the original potential energy hypersurface in such a way as to obtain only a single minimum which, in most cases, is related to the global one. This single minimum can easily be attained from any starting point of the modified hypersurface by standard local minimization procedures. The position of this minimum with respect to the global one in the original hypersurface may have been changed during deformation; therefore, a reversing procedure is applied in which the global minimum is usually attained by gradually reversing the deformation. The hypersurface is deformed with the aid of the diffusion (or heat conduction) equation, with the original shape of the hypersurface having a meaning analogous to the initial concentration (or temperature) distribution. The theory for the development of the DEM is given in Ref. [60]; more recently [38], the diffusion equation was derived from the Schrödinger equation.

The DEM has been applied to a variety of simple mathematical functions [60], and to a series of clusters of Lennard–Jones particles [61]. In the latter application, the Lennard–Jones potential function was expressed as a sum of Gaussians. Calculations were carried out for various cluster sizes $n = 5$, 6,

7,..., 55. For $n = 55$, there are $\approx 10^{45}$ local minima, the global minimum being the Mackay icosahedron. The global minimum was found by the diffusion equation method [61] in ca. 400 s on one processor of an IBM 3090 supercomputer. The DEM has also been applied to water clusters using the MCY potential to treat water–water interactions [63].

We also used the DEM to treat our ECEPP potential, and applied it to terminally blocked alanine and to the pentapeptide Met-enkephalin [62]. The DEM found the global minimum for the alanine compound in < 1 min and a structure close to the global minimum for the pentapeptide in ca. 10 min, using one processor of an IBM 3090 supercomputer. We solved the diffusion equation in the space of Cartesian coordinates of all the atoms of the molecule, and then minimized the solution on the manifold of fixed bond lengths and bond angles, as described in Ref. [62]. This approximation is not entirely satisfactory for minimizing an energy function that is expressed as a sum of pairwise interactions. Each atom–atom pairwise potential is transformed by the diffusion equation into a function that varies as r^2 for large times and in the range of interatomic distances r in a molecule. Of course, such an harmonic function has only one minimum for clusters (a collapsed configuration, i.e., $r = 0$) but many minima for constrained problems, i.e., for molecules. Hence, we are implementing an improvement in the DEM in order to assure that only one minimum remains in the deformed surface at t_0, the time at which the reversing procedure is started. This improvement is necessary because, unlike the calculations on clusters of free Lennard–Jones particles [61], proper connectivity of the chain must be maintained when dealing with polypeptides.

Therefore, for two purposes [(i) maintaining proper connectivity, and (ii) assuring that only one minimum remains in the deformed surface at t_0], we are testing a modification of the DEM by averaging the energy function over conformations that are close to the manifold of fixed bond lengths and bond angles. We construct a linear subspace which is tangent to the manifold at a certain conformation, and average the energy function in this subspace. This procedure is valid at small values of t. For large values of t we use a different method, based on a Fourier expansion of the energy as a function of the dihedral angles; among other things, this involves use of the end-to-end distribution function for a finite freely rotating chain [97]. This modification of the DEM is currently being implemented for large polypeptides [38]; initial tests on small peptides (Met-enkephalin) show that the new approach finds better minima than the previous one.

In another application of the DEM, the procedure is being used to compute crystal structures of small molecules [98].

11. Self-consistent multitorsional field (SCMTF) method

The SCMTF method [64] is based on the idea that the ground-state wave function ψ of a system of nuclei in a molecule spreads over the entire potential energy surface, irrespective of the number of potential wells. Moreover, the maximum of $|\psi|^2$ should lie close to the global minimum of the potential energy. The method treats the dihedral angles, θ_i, of the polypeptide as independent and, by application of the variational principle, leads to a set of N coupled one-dimensional Schrödinger equations, one for each θ_i, to obtain the corresponding $\phi_i(\theta_i)$, where

$$\psi = \prod_i \phi_i(\theta_i)$$

and N is the number of dihedral angles.

$$\hat{H}_i \phi_i = \varepsilon_i \phi_i \quad (i = 1,...,N) \tag{1}$$

with

$$\hat{H}_i = \hat{T}_i + \hat{V}_i^{\text{eff}}(\theta_i) \tag{2}$$

$\hat{T}_i$ is the kinetic energy operator, and $\hat{V}_i^{\text{eff}}$ is the potential energy operator. Each equation describes the variations of a single dihedral angle in the average field of the others. The Hamiltonian of Eq. 1 is

$$\hat{H}_i = \frac{-\hbar^2}{2I_i}\frac{\partial^2}{\partial\theta_i^2} + \hat{V}_i^{\text{eff}}(\theta_i) \tag{3}$$

where I_i is a moment of inertia, and the effective potential $\hat{V}_i^{\text{eff}}(\theta_i)$ depends on the mean field created by averaging over the other dihedral angles, θ_ℓ ($\ell \neq i$), according to the probability density distribution

$\rho_\ell^o = |\phi_\ell^o|^2$. In order to calculate $\hat{V}_i^{\text{eff}}$, a Monte Carlo procedure is used, i.e.,

$$\hat{V}_i^{\text{eff}}(\theta_i) \cong \frac{1}{M^c} \sum_m^{M^c} V(\theta_1^{m*},...,\theta_{i-1}^{m*},\theta_i,\theta_{i+1}^{m*},...,\theta_N^{m*}) \tag{4}$$

where the summation extends over M^c locally minimized trial points in the $(N-1)$-dimensional space of all dihedral angles θ_ℓ except θ_i. First, each point θ^m in the space is selected randomly according to some preassumed one-dimensional distribution ρ_ℓ^o. Then, whenever a θ^m is chosen, the potential energy $V(\theta)$ is minimized with respect to all θ's. The minimization gives the new θ^{m*}, which is then used in Eq. 4. Solving Eq. 1 gives a new set of ϕ_ℓ^o and, therefore, a new set of ρ_ℓ^o. The procedure is repeated iteratively until self-consistency of the ρ_ℓ^o distributions is achieved. The ECEPP potential is used for $V(\theta)$. The SCMTF method has been applied to Met-enkephalin [64], icosalanine [65], and the 20-residue membrane-bound portion of melittin [66].

12. Minimization-in-a-confined-space approach (CSA)

Since hydrophobic interactions lead to compact structures, we have imposed the restraint of confining the polypeptide chain to fold within a compact volume [67–69]. On the basis of these calculations, we have concluded [67] that the observed conformations of native proteins may arise from two basic factors: the compactness of structures under hydrophobic interactions and the intrinsic stiffness of the polypeptide chains due to the interactions within terminally blocked residues. An outgrowth of this work is the confined-space approach (CSA) which is being used to examine unfolding and folding pathways.

In the CSA, the unfolding and refolding of the native structure of a protein (BPTI in the example examined) are characterized by the dimensions of the protein, expressed in terms of the three principal radii, r_a, r_b, r_c, of the structure considered as an ellipsoid [68]. A dynamic equation, describing the variations of the principal radii on the unfolding path, and a numerical procedure to solve this equation, are used. In order to refold the expanded and distorted conformations to the native structure, a dimensionally constrained function is introduced in a minimization procedure. A unique and reproducible unfolding pathway for an intermediate of BPTI lacking the [30,51] disulfide bond was obtained. (If this disulfide bond is intact, native BPTI cannot be unfolded unless the procedure is given an unreasonably high initial kinetic energy.) The most interesting finding is that the majority of expanded conformations, generated by the dynamic procedure under various conditions, can be refolded closely to the native structure, as measured by the correct overall chain fold, by the r.m.s. deviations from the native structure of only 1.9–3.1 Å, and by the energy differences of about 10 kcal/mol from the native structure. Introduction of the [30,51] disulfide bond at this stage, followed by minimization, improved the closeness of the refolded structures to the native structure, reducing the r.m.s. deviations to 0.9–2.0 Å.

The longest principal radii of the expanded conformations are more than 2.5 times greater than those of the native structure. This increase in the size of the molecule is much larger than the 10–25% that is usually obtained by ordinary molecular dynamics in the available computer time of this procedure. The correct refolding of our unfolded structures by encompassing such a large conformational space indicates that these expanded conformations have a strong tendency to move toward the native structure. The results indicate that the folding of these expanded conformations, under conditions that favor the compact native structure as mimicked by the volume constraints in our algorithm, is strongly dictated by internal interactions in the amino acid sequence of BPTI and that folding follows some unique pathways, as suggested by the fact that several refolding simulations follow a similar path. These results have implications for the general protein folding problem, for the effects of volume constraints (simulating the collapsing role of hydrophobic interactions, as shown for lattice models by Dill [12,15], Shakhnovich [16,17,19,27,33], Skolnick [20,22,23], Hao and Scheraga [34–36], and others) on the formation of the native structures of proteins, and for the general applicability of this procedure.

The unfolding is carried out by solving the dynamic equation

$$[\mathrm{M}]\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -\frac{\mathrm{d}U(x)}{\mathrm{d}x} \tag{5}$$

where [M] is a diagonal mass matrix, x is the transpose of the vector $[r_a, r_b, r_c]$, and $U(x)$ is the potential energy of the protein molecule; the solvent effect is included implicitly in the kinetic energy. This differential equation is solved by the backward-Euler scheme [99] involving a minimization method proposed by Schlick [99]. It is this minimization which takes this procedure into an experimentally reasonable time scale, in contrast to that in ordinary molecular dynamics. The unfolded protein is refolded by minimizing in a confined space, i.e., the following function is minimized:

$$F = U + K\left[(r_a - \rho_a)^2 + (r_b - \rho_b)^2 + (r_c - \rho_c)^2\right] \tag{6}$$

where K is the weight for the constraint term, and ρ_a, ρ_b, ρ_c are the target values of r_a, r_b, r_c, taken from the crystal structure. Details of the computational method, the results, and the computer timings, are given in Ref. [68].

The calculations of Ref. [68] used the values of ρ_a, ρ_b, ρ_c taken from the crystal structure. It may be possible to assign these parameters without knowing the crystal structure in advance. For this purpose, it may be possible to use the observations [67] that the volume of a protein (from the Brookhaven Protein Data Bank) is a linear function of the number n of residues, and that the eccentricities of the ellipsoidal shapes of these proteins is independent of n, within a certain tolerance. Therefore, given only a knowledge of n of an unknown protein, one can compute the parameters ρ_a, ρ_b, ρ_c. Since there is some tolerance in these values, because of the observed tolerance in the eccentricity, it would be necessary to experiment with procedures to include these tolerances by including deviations in ρ_a, ρ_b, ρ_c, with various trial-and-error weights assigned to these deviations.

It would also be worthwhile to explore the influence of certain approximations that were made in Ref. [68], viz., we previously considered only stretching and compressing deformations but neglected the shearing components of the deformation in evaluating the kinetic energy of unfolding; also, we assumed a uniform deformation of the protein structure along each coordinate direction. It may be possible to dispense with these approximations.

13. Lattice neural network minimization (LNNM)

A rapid lattice neural network minimization (LNNM) procedure has been used to locate the global-minima conformations of proteins [70]. The conformation of a protein is represented as an array of amino acid sequence vs. position on a three-dimensional face-centered cubic lattice with an energy function defined in terms of the array variables. Using the LNNM method, the energy function is minimized to locate the global minimum energy for the conformation of the protein. The energy function consisted of site exclusion and bond connectivity penalty terms and a pairwise contact energy potential. The contact energy potential used in the procedure is the united-residue potential of Miyazawa and Jernigan [100]. The LNNM method found the global minimum for a seven-residue peptide in all of the 15 runs carried out. For a nine-residue peptide, the global minimum was found in 7 out of 15 runs, and the global minimum or the second lowest minimum in 10 of the runs; in contrast a Monte Carlo simulated annealing method found the global minimum or the second lowest minimum in only two runs, in the same total CPU time.

Starting from a uniform array on the lattice for the 46-residue protein, crambin, the energy of the crambin array was minimized and a compact low-energy structure was found in ca. 25 min of CPU time on one processor of an IBM 3090 computer. Its energy was much lower than that of the native protein, suggesting that there are inadequacies in the potential of Miyazawa and Jernigan.

The LNNM method was also applied to predict chain-folding initiation sites (CFIS) [101,102] of a protein. The LNNM method correctly predicted the CFIS for the two proteins examined, RNase S and T4 lysoyzme. The method was also applied to another chain optimization problem, minimization of the r.m.s. distance error in fitting X-ray structures to a lattice, with good results.

14. Concluding remarks

With the aid of lattice models, analytical theories, and computer simulation, progress is being made toward an understanding of the statistical mechanical aspects of the protein folding problem and of the interatomic interactions that determine how a polypeptide chain folds into the unique three-dimensional structure of a native protein. Similarly, progress is being made toward surmounting the multiple-minima problem. This problem has already been solved for small open-chain and cyclic peptides, and for regular-repeating models of fibrous proteins; methods are being developed to solve this problem for globular proteins. Solution of this problem will make it feasible to use energy-minimization and Monte Carlo calculations to identify the global minimum in the multi-dimensional energy landscape of a protein. Some of the methods developed here are applicable to other global optimization problems in physical chemistry, e.g. the computation of crystal structures.

Acknowledgements

This work has been supported by grants from the National Institutes of Health (GM-14312) and from the National Science Foundation (DMB90-15815).

References

[1] C.B. Anfinsen, E. Haber, M. Sela and F.H. White, Jr., Proc. Natl. Acad. Sci. USA, 47 (1961) 1309–1314.
[2] G. Némethy and H.A. Scheraga, Biopolymers, 3 (1965) 155–184.
[3] G.N. Ramachandran, D. Ramakrishnan and V. Sasisekharan, J. Mol. Biol., 7 (1963) 95–99.
[4] J.A. Schellman and C. Schellman, in H. Neurath (Ed.), The Proteins: Composition, Structure and Function, 2nd edn., Academic Press, New York, 1964, pp. 1–137.
[5] D.C. Poland and H.A. Scheraga, Biopolymers, 3 (1965) 275–419.
[6] D. Poland and H.A. Scheraga, in G.D. Fasman (Ed.), Poly-α-Amino Acids, Marcel Dekker, New York, 1967, pp. 391–497.
[7] H.A. Scheraga, Adv. Phys. Org. Chem., 6 (1968) 103–184.
[8] K.D. Gibson and H.A. Scheraga, Physiol. Chem. Phys., 1 (1969) 109–126.
[9] N. Go and H.A. Scheraga, J. Chem. Phys., 51 (1969) 4751–4767.
[10] N. Go and H.A. Scheraga, Macromolecules, 9 (1976) 535–542.
[11] H. Taketomi, Y. Ueda and N. Go, Int. J. Peptide Protein Res., 7 (1975) 445–459.
[12] K.A. Dill, Biochemistry, 24 (1985) 1501–1509.
[13] J.D. Bryngelsen and P.G. Wolynes, Proc. Natl. Acad. Sci. USA, 84 (1987) 7524–7528.
[14] J.D. Bryngelsen and P.G. Wolynes, J. Phys. Chem., 93 (1989) 6902–6915.
[15] K.A. Dill, D.O.V. Alonso and K. Hutchinson, Biochemistry, 28 (1989) 5439–5449.
[16] E.I. Shakhnovich and A.V. Finkelstein, Biopolymers, 28 (1989) 1667–1680.
[17] E.I. Shakhnovich and A.M. Gutin, Biophys. Chem., 34 (1989) 187–199.
[18] D.G. Covell and R.L. Jernigan, Biochemistry, 29 (1990) 3287–3294.
[19] E.I. Shakhnovich and A.M. Gutin, Nature, 346 (1990) 773–775.
[20] J. Skolnick and A. Kolinski, Science, 250 (1990) 1121–1125.
[21] J.D. Honeycutt and D. Thirumalai, Proc. Natl. Acad. Sci. USA, 87 (1990) 3526–3529.
[22] J. Skolnick and A. Kolinski, J. Mol. Biol., 221 (1991) 499–531.
[23] A. Kolinski and J. Skolnick, J. Chem. Phys., 97 (1992) 9412–9426.
[24] D.A. Hinds and M. Levitt, Proc. Natl. Acad. Sci. USA, 89 (1992) 2539–2540.
[25] P.E. Leopold, M. Montal and J.N. Onuchic, Proc. Natl. Acad. Sci. USA, 89 (1992) 8721–8725.
[26] C.J. Comacho and D. Thirumalai, Proc. Nat. Acad. Sci. USA, 90 (1993) 6369–6372.
[27] E.I. Shakhnovich and A.M. Gutin, Proc. Natl. Acad. Sci. USA, 90 (1993) 7195–7199.
[28] M. Fukugita, D. Lancaster and M.G. Mitchard, Proc. Natl. Acad. Sci. USA, 90 (1993) 6365–6368.
[29] A. Kolinski, A. Godzik and J. Skolnick, J. Chem. Phys., 98 (1993) 7420–7433.
[30] A. Kolinski and J. Skolnick, Proteins, 18 (1994) 338–352.
[31] A. Kolinski and J. Skolnick, Proteins, 18 (1994) 353–366.
[32] M. Vieth, A. Kolinski, C.L. Brooks III and J. Skolnick, J. Mol. Biol., 237 (1994) 361–367.
[33] A. Sali, E.I. Shakhnovich and M. Karplus, Nature, 369 (1994) 248–251.
[34] M.H. Hao and H.A. Scheraga, J. Phys. Chem., 98 (1994) 4940–4948.
[35] M.H. Hao and H.A. Scheraga, J. Phys. Chem., 98 (1994) 9882–9893.
[36] M.H. Hao and H.A. Scheraga, J. Chem. Phys., 102 (1995) 1334–1348.
[37] H.A. Scheraga, M.H. Hao and J. Kostrowicki, in M.Z. Atassi (Ed.), Methods in Protein Structure Analysis, Plenum Press, in press.
[38] J. Kostrowicki and H.A. Scheraga, in P. Pardalos, D. Shal-

loway and G. Xue (Eds.), Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, 1995.
[39] M.H. Hao and H.A. Scheraga, 1995 Supercomputer Symposium, in press.
[40] H.A. Scheraga, in K.B. Lipkowitz and D.B. Boyd (Eds.), Reviews in Computational Chemistry, Vol. 3, VCH Publishers, New York, 1992, pp. 73–142.
[41] K.D. Gibson and H.A. Scheraga, in R.H. Sarma and M.H. Sarma (Eds.), Structure & Expression, Vol. 1, From Proteins to Ribosomes, Adenine Press, Guilderland, NY, 1988, pp. 67–94.
[42] M. Vásquez, G. Némethy and H.A. Scheraga, Chem. Revs., 94 (1994) 2183–2239.
[43] H.A. Scheraga, in C.B. Anfinsen and A.N. Schechter (Eds.), Current Topics in Biochemistry, 1973, Academic Press, New York, 1974, pp. 1–42.
[44] I. Simon, G. Némethy and H.A. Scheraga, Macromolecules, 11 (1978) 797–804.
[45] M. Vasquez and H.A. Scheraga, Biopolymers, 24 (1985) 1437–1447.
[46] L. Piela and H.A. Scheraga, Biopolymers, 26 (1987) S33–S58.
[47] Z. Li and H.A. Scheraga, Proc. Natl. Acad. Sci. USA, 84 (1987) 6611–6615.
[48] Z. Li and H.A. Scheraga, J. Mol. Struct. (Theochem), 179 (1988) 333–352.
[49] D.R. Ripoll and H.A. Scheraga, Biopolymers 27 (1988) 1283–1303.
[50] D.R. Ripoll and H.A. Scheraga, J. Protein Chem., 8 (1989) 263–287.
[51] D.R. Ripoll and H.A. Scheraga, Biopolymers, 30 (1990) 165–176.
[52] D.R. Ripoll, M.J. Vasquez and H.A. Scheraga, Biopolymers, 31 (1991) 319–330.
[53] D.R. Ripoll, L. Piela, M. Vasquez and H.A. Scheraga, Proteins: Struct. Funct. Genet., 10 (1991) 188–198.
[54] G.H. Paine and H.A. Scheraga, Biopolymers, 24 (1985) 1391–1436.
[55] G.H. Paine and H.A. Scheraga, Biopolymers, 25 (1986) 1547–1563.
[56] G.H. Paine and H.A. Scheraga, Biopolymers, 26 (1987) 1125–1162.
[57] E.O. Purisima and H.A. Scheraga, Proc. Natl. Acad. Sci. USA, 83 (1986) 2782–2786.
[58] E.O. Purisima and H.A. Scheraga, J. Mol. Biol., 196 (1987) 697–709.
[59] M.H. Lambert and H.A. Scheraga, J. Comput. Chem., 10 (1989) 770–797, 798–816, 817–831.
[60] L. Piela, J. Kostrowicki and H.A. Scheraga, J. Phys. Chem., 93 (1989) 3339–3346.
[61] J. Kostrowicki, L. Piela, B.J. Cherayil and H.A. Scheraga, J. Phys. Chem., 95 (1991) 4113–4119.
[62] J. Kostrowicki and H.A. Scheraga, J. Phys. Chem., 96 (1992) 7442–7449.
[63] R.J. Wawak, M.M. Wimmer and H.A. Scheraga, J. Phys. Chem., 96 (1992) 5138–5145.
[64] K.A. Olszewski, L. Piela and H.A. Scheraga, J. Phys. Chem., 96 (1992) 4672–4676.
[65] K.A. Olszewski, L. Piela and H.A. Scheraga, J. Phys. Chem., 97 (1993) 260–266.
[66] K.A. Olszewski, L. Piela and H.A. Scheraga, J. Phys. Chem., 97 (1993) 267–270.
[67] M.H. Hao, S. Rackovsky, A. Liwo, M.R. Pincus and H.A. Scheraga, Proc. Natl. Acad. Sci. USA, 89 (1992) 6614–6618.
[68] M.H. Hao, M.R. Pincus, S. Rackovsky and H.A. Scheraga, Biochemistry, 32 (1993) 9614–9631.
[69] A. Liwo, M.R. Pincus, R.J. Wawak, S. Rackovsky and H.A. Scheraga, Protein Sci., 2 (1993) 1715–1731.
[70] A.A. Rabow and H.A. Scheraga, J. Mol. Biol., 232 (1993) 1157–1168.
[71] S. Kirkpatrick, C.D. Galatt, Jr. and M.P. Vecchi, Science, 220, (1983) 671–680.
[72] D. Vanderbilt and S.G. Louie, J. Comput. Phys., 56 (1984) 259–271.
[73] M. Nilges, A.M. Gronenborn, A.T. Brunger and G.M. Clore, Protein Eng., 2 (1988) 27–38.
[74] A. Nayeem, J. Vila and H.A. Scheraga, J. Comput. Chem., 12 (1991) 594–605.
[75] L. Glasser and H.A. Scheraga, J. Mol. Biol., 199 (1988) 513–524.
[76] M. Dygert, N. Go and H.A. Scheraga, Macromolecules, 8 (1975) 750–761.
[77] G. Némethy and H.A. Scheraga, Biochem. Biophys. Res. Commun., 118 (1984) 643–647.
[78] S.E. Hull, R. Karlsson, P. Main, M.M. Woolfson and E.J. Dodson, Nature, 275 (1978) 206–207.
[79] P.A. Mirau and F.A. Bovey, Abstracts of the 199th Am. Chem. Soc., Boston, MA, April, 1990, p. POLY 58.
[80] Y. Xu, I.P. Sugar and N.R. Krishna, J. Biomolec. NMR, 5 (1995) 37–48.
[81] M.H. Miller and H.A. Scheraga, J. Polym. Sci. Polym. Symp., 54 (1976) 171–200.
[82] M.H. Miller, G. Némethy and H.A. Scheraga, Macromolecules, 13 (1980) 470–478, 910–913, 914–919.
[83] K. Okuyama, N. Tanaka, T. Ashida and M. Kakudo, Bull. Chem. Soc. Jpn., 49 (1976) 1805–1810.
[84] S. Fossey, G. Némethy, K.D. Gibson and H.A. Scheraga, Biopolymers, 31 (1991) 1529–1541.
[85] M. Vasquez and H.A. Scheraga, J. Biomol. Struct. Dyn., 5 (1988) 705–755, 757–784.
[86] M.J. Sippl and H.A. Scheraga, Proc. Natl. Acad. Sci. USA, 83 (1986) 2283–2287.
[87] G.M. Crippen, Distance Geometry and Conformational Calculations, Research Studies Press, Chichester, 1981.
[88] W. Braun, C. Bösch, L.R. Brown, N. Go and K. Wüthrich, Biochim. Biophys. Acta, 667 (1981) 377–396.

[89] F.A. Momany, R.F. McGuire, A.W. Burgess and H.A. Scheraga, J. Phys. Chem., 79 (1975) 2361–2381.
[90] G. Némethy, M.S. Pottle and H.A. Scheraga, J. Phys. Chem., 87 (1983) 1883–1887.
[91] M.J. Sippl, G. Némethy and H.A. Scheraga, J. Phys. Chem., 88 (1984) 6231–6233.
[92] G. Némethy, K.D. Gibson, K.A. Palmer, C.N. Yoon, G. Paterlini, A. Zagari, S. Rumsey and H.A. Scheraga, J. Phys. Chem., 96 (1992) 6472–6484.
[93] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, Jr., M.D. Bricc, J.R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, J. Mol. Biol., 112 (1977) 535–542.
[94] A. Kidera, Y. Konishi, M. Oka, T. Ooi and H.A. Scheraga, J. Protein Chem., 4 (1985) 23–55.
[95] T.L. Blundell, J.E. Pitts, I.J. Tickle, S.P. Wood and C.W. Wu, Proc. Natl. Acad. Sci. USA, 78 (1981) 4175–4179.
[96] I. Glover, I. Haneef, J. Pitts, S. Wood, D. Moss, I. Tickle and T. Blundell, Biopolymers, 22 (1983) 293–304.
[97] J. Kostrowicki and H.A. Scheraga, Comput. Polym. Sci., 5 (1995) 47–55.
[98] R.J. Wawak, K.D. Gibson, A. Liwo and H.A. Scheraga, Proc. Natl. Acad. Sci. USA, in press.
[99] T. Schlick and W.K. Olson, J. Mol. Biol., 223 (1992) 1089–1119.
[100] S. Miyazawa and R.L. Jernigan, Macromolecules, 18 (1985) 534–552.
[101] R.R. Matheson, Jr. and H.A. Scheraga, Macromolecules, 11 (1978) 819–829.
[102] G.T. Montelione and H.A. Scheraga, Acc. Chem. Res., 22 (1989) 70–76.