[Frontiers in Bioscience 9, 3296-3323, September 1, 2004]

THE PROTEIN FOLDING PROBLEM: GLOBAL OPTIMIZATION OF FORCE FIELDS

H. A. Scheraga, A. Liwo, S. Oldziej, C. Czaplewski, J. Pillardy, D. R. Ripoll, J. A. Vila, R. Kazmierkiewicz, J. A. Saunders, Y. A. Arnautova, A. Jagielska, M. Chinchio, and M. Nanias

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

TABLE OF CONTENTS

1. ABSTRACT

The evolutionary development of a theoretical approach to the protein folding problem, in our laboratory, is traced. The theoretical foundations and the development of a suitable empirical all-atom potential energy function and a global optimization search are examined. Whereas the all-atom approach has thus far succeeded for relatively small molecules and for α-helical proteins containing up to 46 residues, it has been necessary to develop a hierarchical approach to treat larger proteins. In the hierarchical approach to single- and multiple-chain proteins, global optimization is carried out for a simplified united residue (UNRES) description of a polypeptide chain to locate the *region* in which the global minimum lies. Conversion of the UNRES structures in this region to all-atom structures is followed by a local search in this region. The performance of this approach in successive CASP blind tests for predicting protein structure by an ab initio physics-based method is described. Finally, a recent attempt to compute a folding pathway is discussed.

2. INTRODUCTION

Concern about the protein folding problem essentially started with the famous experiment of Anfinsen (1) who demonstrated that unfolded bovine pancreatic ribonuclease A (RNase A), with its four disulfide bonds reduced, could be refolded spontaneously by oxidation of all of its sulfhydryl groups to re-form the native, biologically-active conformation. This experiment formed the basis of the thermodynamic hypothesis, wherein the native conformation is thought to be the thermodynamically most stable one, i.e., the one with the lowest free energy of the system (protein plus solvent). This hypothesis implies that the amino acid sequence of the protein contains all the information required for proper folding, and has guided theoretical computational efforts to compute the native conformation from a knowledge of only the amino acid sequence.

There are really two protein folding problems. The first is to compute the three-dimensional structure of the native protein based only on the thermodynamic hypothesis and the amino acid sequence. The second is to compute the structural pathways and rates by which the completely unfolded polypeptide chain proceeds to the folded native conformation. Most of the research that followed from Anfinsen's experiment has been concerned with the first problem. In recent years, efforts have been made to compute folding pathways. However, the question of folding pathways has also been explored experimentally, frequently using Bovine Pancreatic Trypsin Inhibitor (BPTI) (2, 3) or RNase A (4) as the protein for study.

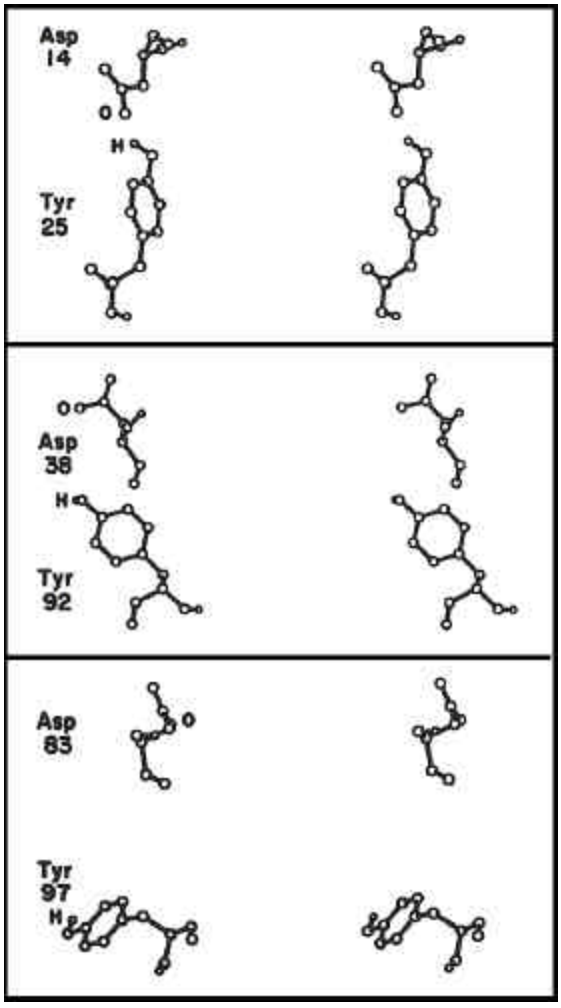

Figure 1. Verification of three predicted tyrosyl•••aspartate interactions (5) by the subsequently-determined x-ray structure of RNase A.

There are experimental x-ray and NMR approaches to determine three-dimensional structures of proteins. There are also a variety of spectroscopic and kinetic methods to determine folding pathways. So, what is the point of introducing the computational approach? The answer is clearly that the computational approach can provide an understanding of how the physics of inter-residue interactions leads to the final folded structure and to the pathways to reach the native structure.

The computational work from our own laboratory, which forms the basis of this article, had its inception from our experimental studies of RNase A. With physical chemical and spectroscopic studies, we obtained several distance constraints (5) that must be satisfied by the folded structure. These are shown in Figure 1 (6) and in Figures 2-4. This motivated our development of computational procedures (7-9) to identify the folded structure, using these distance constraints and, ultimately, to compute protein structure solely from the physics of the inter-residue interactions even without reliance on experimentally-determined distance constraints.

We also used RNase A to explore the folding pathways of this protein experimentally (4), and recently began to approach this problem theoretically (10). This article focuses on the theoretical approach.

3. INGREDIENTS OF THE COMPUTATIONAL PROCEDURE

There are two basically different approaches to compute protein structure, — a knowledge-based one and a physics-based ab initio approach. The knowledge-based approach makes use of information about protein structures that have already been solved by x-ray or NMR methods, such as secondary-structure prediction, homology modeling, threading, or fragment coupling. The ab initio approach, based on the physics of the inter-residue interactions, makes no use of knowledge-based information in the search for the thermodynamically most stable state. Thus far, the knowledge-based approach has been more successful than the ab initio one, but the latter is catching up. More important, it is only the ab initio approach that can provide an understanding of how physics governs the folding pathways and the final structure.

The two essential ingredients of the computational procedure are the force field and the search procedure. The force field consists of the potential energy and entropy contributions (11-13). For proteins, various empirical potential energy functions have been proposed, e.g., ECEPP (14-17), AMBER (18), CHARMM (19), DISCOVER (20), including explicit (21, 22) and implicit (23-28) treatment of the solvent. Two sources of entropy are considered, viz., the conformational entropy [computed from the matrix of second derivatives of the energy in a harmonic approximation (11, 12)], and the free energy of solvation when using implicit solvent models (23-28). For explicit treatment of hydration, the potential energy is used directly. The search procedure focuses on identifying the global minimum of the potential energy (9). The overall procedure is considered to be an ab initio one in the sense that no knowledge-based information is used in the search procedure, even though knowledge-based information is sometimes used to obtain the potential energy. However, more and more of the knowledge-based potential functions are being replaced by quantum mechanical calculations on model compounds (29-31). Most of the attention in this article is focused on consideration of global optimization search procedures.

4. MAGNITUDE OF THE PROBLEM

Figure 5 is a representation of a portion of the polypeptide chain. Our force field, ECEPP (Empirical Conformational Energy Program for Peptides), keeps the bond lengths and bond angles fixed, and varies only the dihedral angles for rotation about backbone and side-chain bonds (14-17). Because of the partial double bond character of the C'-N peptide bond, its rotation is very restricted, leading to small variations about the cis and trans

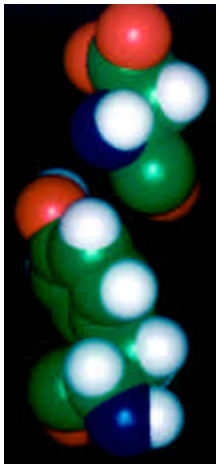

Figure 2. Space-filling model of Tyr 25•••Asp 14 interaction of Figure 1.

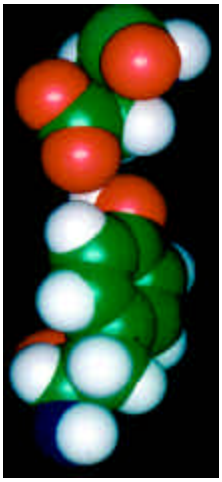

Figure 3. Space-filling model of Tyr 92•••Asp 38 interaction of Figure. 1.

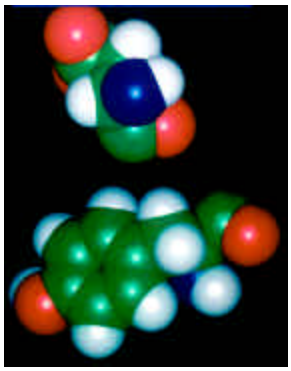

Figure 4. Space-filling model of Tyr 97•••Asp 83 interaction of Figure 1.

conformations, respectively. Other force fields, such as AMBER (18) CHARMM (19), and DISCOVER (20) allow for variations of bond lengths and bond angles (32). As shown in Figure 5, there are three backbone dihedral angles and an average of three side-chain dihedral angles, or a total of six degrees of internal rotational freedom for each amino acid residue. Therefore, there are 600 degrees of freedom for a 100-residue polypeptide chain if the bond lengths and bond angles are fixed. If the bond lengths and bond angles are allowed to vary, there is a large increase in the number of degrees of freedom. It is thus clear that the potential energy surface is a complex multi-dimensional one on which, according to the thermodynamic hypothesis, one must search for the lowest or global minimum on this surface. In separate articles (9, 33), we have summarized the various global optimization approaches used in these computations. In our laboratory, we have focused on minimization and Monte Carlo procedures. We generally avoid molecular dynamic procedures for ab initio folding with all-atom models because the required femtosecond time step cannot reach the millisecond-to-second time scale for the folding of most proteins with presently available computer resources.

5. APPLICATION OF SOME OF THE GLOBAL OPTIMIZATION APPROACHES

5.1. Build-up

In the build-up procedure (34, 37), use is made of the whole potential energy (f, y) [Ramachandran (38)] map of each residue. There are of the order of 10 local minima on each map. A dipeptide is then built from two residues taking the 10 x 10 combinations of the local minima of each map as the initial points, for 100 energy minimizations. The resulting minima pertain to slightly perturbed conformations from the starting single residues because of inter-residue interactions. The chain is thus built up in this manner by adding the local minima from each successive residue to the growing chain, followed by energy minimization at each stage. As more and more residues are added, increasing contributions from long-range interactions come into play, and the set of long-range interactions becomes complete when the last residue is added to the chain. To save computation time, the very high-energy conformations are discarded at various stages of the build-up procedure.

The build-up procedure has been applied to a variety of linear peptides, including the membrane-bound portion of melittin (35) and the pentapeptide methionine enkephalin (36). However, the global-minimum structure of enkephalin, shown in Figure 6, was not attained until after the development of the MCM procedure (see section 5.2) and later procedures (see, e.g., section 5.4).

Another example of the application of the build-up procedure is the cyclic decapeptide gramicidin S (39). Because the molecule is cyclic, the restraint to close the ring exactly (40), with the observed restraint of C_2 symmetry (41), was added to the build-up procedure. The resulting minimum-energy structure is shown in Figure 7. It was verified by a multi-dimensional NMR study by Mirau and Bovey (42), who used the published coordinates (39) to compute the NMR spectrum and stated, "we compare the

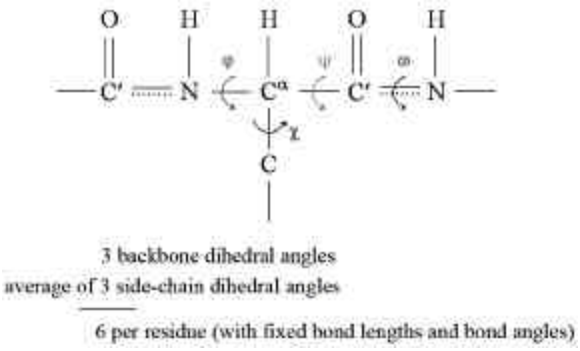

Figure 5. Representation of a portion of the polypeptide chain showing the backbone and side-chain dihedral-angle degrees of freedom to alter the conformation, keeping the bond lengths and bond angles fixed as in the ECEPP force field.

Figure 6. Lowest-energy computed conformation of the pentapeptide methionine-enkephalin, obtained with the ECEPP force field.

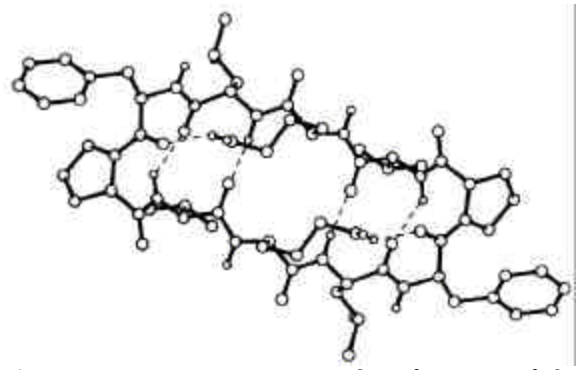

Figure 7. Lowest-energy computed conformation of the cyclic decapeptide gramicidin S.

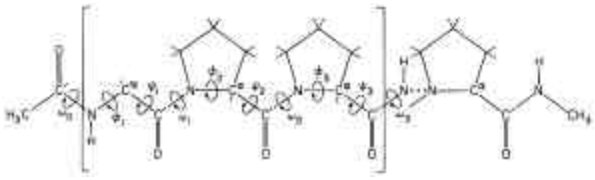

Figure 8. Representation of the polytripeptide, Poly(Gly-Pro-Pro) model of collagen.

experimental spectrum of gramicidin S with the theoretical spectrum calculated from the atomic coordinates of the energy-minimized structure of Scheraga and coworkers. Close agreement is obtained for the backbone protons".

As an additional example of the build-up procedure, we cite its application to collagen-like polytripeptides. In collagen, every third residue is glycine, and the other two residues of the tripeptide are frequently proline or hydroxyproline, as illustrated in Figure 8 for poly (Gly-Pro-Pro). In the build-up procedure, the energies of the dipeptides Gly-Pro, Pro-Pro, and Pro-Gly, and then the tripeptide Gly-Pro-Pro, were calculated. Single chains were then built from repeating tripeptides, and the resulting chains were then packed in different symmetries for the (experimentally-observed) three-chain character of the complex. The resulting minimum-energy coiled-coil structure (43) is shown in Figure 9. It agrees with the limited parameters for the coiled-structure obtained from fiber diffraction experiments on natural collagen by Ramachandran and Kartha (44), Rich and Crick (45), and Yonath and Traub (46). Subsequent single-crystal studies of $(Gly\text{-}Pro\text{-}Pro)_{10}$ carried out in Japan (47), Rutgers (48) and Naples (49) agreed with each other within an rms deviation of about 0.2-0.4 Å, and the computed structure agreed with all of these with an rms deviation of 0.5 Å.

Not all polytripeptides containing glycine in each triad form coiled-coil structures of the type shown in Figure 9. Some, such as poly(Gly-Ala-Pro), tend to form parallel-chain structures (50). Our calculations on these other synthetic poly(tripeptides) (50-52) and also on a natural collagen sequence (53) agree with the observed structures.

Judged by the agreement between experimental and calculated structures, in the above examples, the force field and search procedures are reasonably accurate.

5.2. Monte Carlo-with-Minimization (MCM)

Metropolis Monte Carlo (54) is a poor approach for global optimization of polypeptide chains. It generates a Markov chain which solves the ergodicity problem only with an infinite amount of computer time. However, if Metropolis Monte Carlo is coupled with energy minimization (55, 56), the procedure is more efficient. The algorithm (55) involves the following steps:

1. Select a conformation at random and minimize its ECEPP energy. This leads to a local minimum which is generally not the global minimum.
2. Select any backbone or side-chain dihedral angle at random. In a subsequent version of the procedure (56), more than one dihedral angle is selected.
3. Make random changes in the selected dihedral angles over the whole range (-180° to +180°).
4. Minimize the energy of the new conformation.
5. Compare E_{new} to E_{old} by means of the Metropolis criterion to decide whether to accept E_{new}.
6. Iterate.

In Figure 10, the numbers 1 to 17 on the (f,y) maps of the five residues of methionine enkephalin

Figure 9. Lowest-energy computed conformation of (Gly-Pro-Pro)$_{10}$.

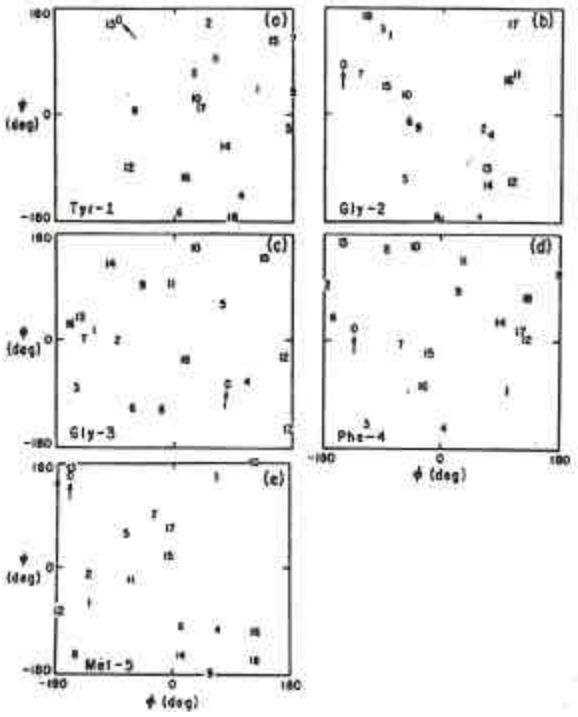

Figure 10. The numbers 1-17 on the five (f, y) maps represent 17 random starting conformations of the residues of methionine enkephalin in the MCM procedure. All 17 independent runs converged to the same global minimum, indicated by the zeros on each map, which is the same as the structure shown in Figure 6.

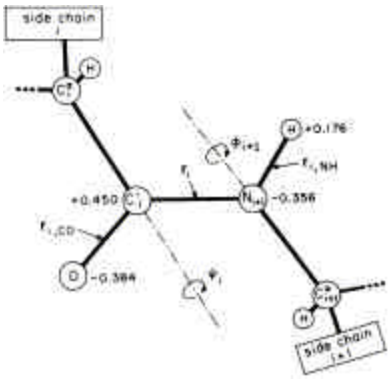

Figure 11. Variation of f, ? to align the peptide-bond dipole.

correspond to 17 randomly selected starting conformations. All 17 runs converged (56) to the conformation indicated by the zeros on each map, and this conformation is identical to that of Figure 6. Thus, several different global optimization procedures (MCM and those cited below) led to the same global minimum with, of course, the same (ECEPP) potential function.

The conformation in Figure 6, however, does not agree with either of two polymorphic crystal structures (57, 58), undoubtedly because these two crystal structures involve intermolecular hydrogen bonds that were not present in the calculations for the single molecule. Therefore, the global-optimization calculations were repeated for three crystal structures, the two observed ones and structures obtained by packing the conformation of Figure 6 in different symmetries (59). The results showed that crystals of the conformation of Figure 6 had higher energy than either of the two minimized experimental structures. However, if the single-molecule conformations in the observed crystal structures are deprived of their intermolecular hydrogen bonds, then the conformation of Figure 6 is indeed lower in energy than either of the conformations in the observed crystal structures. Other global optimization procedures (60-63) also led to the structure of Figure 6, which may therefore be regarded as the global minimum structure of the isolated molecule for the given (ECEPP) potential.

5.3. Self-consistent electric field (SCEF) method

The SCEF method is based on the assumption that the dipole within each peptide group should be optimally aligned in its local electric field (64). This assumption has recently been validated (65). The electric field can be computed at every peptide group for any conformation of the polypeptide chain. From an examination of the alignments of all such dipoles, the worst-aligned is selected and optimally re-aligned by variation of the neighboring values of f and ? (see Figure 11). This procedure regenerated the native structure of an a-helix from a disrupted structure (64), indicating that the orientation of the backbone (peptide-group) dipoles in the local electric field can be used to determine the regions of the polypeptide that are candidates for improvement during conformational searches.

5.4. Electrostatically-driven Monte Carlo (EDMC) method

Additional efficiency is gained by combining MCM with the SCEF method in the so-called EDMC method (61, 66). The EDMC method was applied to several small peptides with good results (67, 68). In its initial application (66), the EDMC method was able to fold a 19-residue chain of poly(L-alanine) into a full a-helix starting from random and arbitrary initial conformations. The procedure even converted an initial left-handed a-helix to its lower-energy right-handed form, surmounting the intervening energy barrier (see Figure 12). It was originally thought to be applicable to chains no longer than 20 residues. However, with the acquisition of a Beowulf-type cluster, illustrated in Figure 13, it has been possible to extend the EDMC method, thus far, to a chain of 46

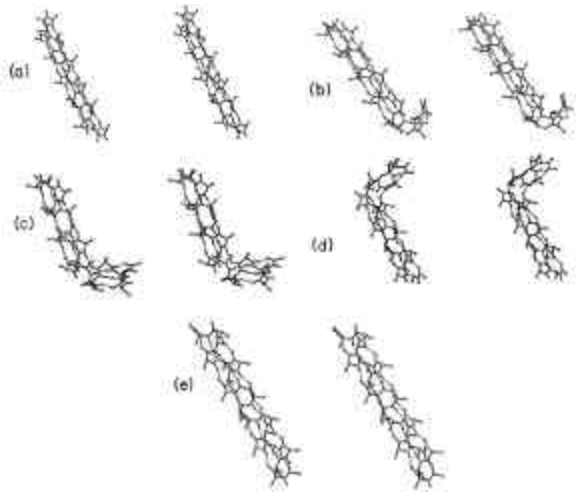

Figure 12. Stereo plots of a set of conformations of poly(L-alanine) encountered on a conformational pathway during an EDMC folding simulation from a left-handed a-helix (a), through intermediate stages (b, c, d), to a right-handed a-helix (e).

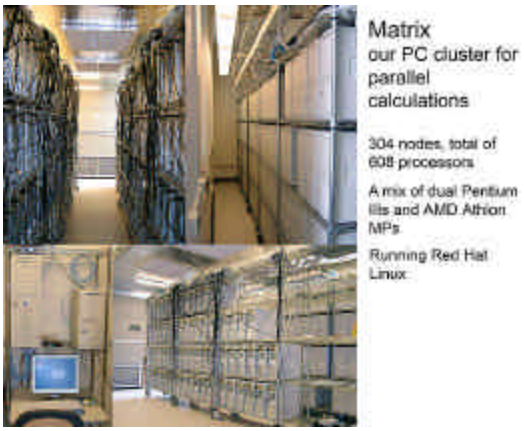

Figure 13. Various views of a Beowulf-type cluster of computers.

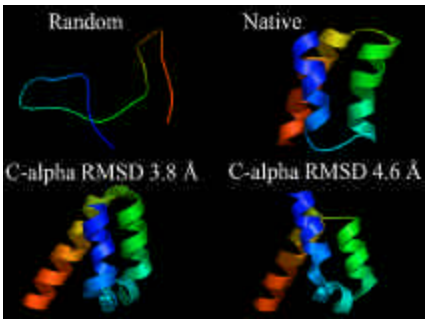

Figure 14. Ribbon diagrams of (a) a random starting conformation of protein A, (b) the native fold obtained by NMR, (c) the lowest-energy structure obtained with the SRFOPT solvation model, and (d) the lowest-energy structure obtained with the OONS solvation model.

residues, the B domain of staphylococcal protein A (69). The EDMC method scales well on Beowulf-type clusters even with slow communication between nodes (70).

Protein A is larger than the 36-residue a-helical protein from the villin headpiece, for which all-atom simulations, starting from an extended structure, were previously carried out (71, 72). Those simulations were carried out with explicit solvent, which increases the computing time considerably compared to the time required for the implicit solvent models used in our simulations (69).

Calculations on protein A were carried out with ECEPP/3 (17) and the EDMC procedure (68), together with two implicit hydration models, OONS (24) and SRFOPT (26), starting from four different random conformations, one of which is illustrated schematically in Figure 14 (69). Three of the four runs converged to the same native-like fold illustrated in Figure 14 for each of the two hydration models; the fourth converged to the mirror-image conformation. A structure even closer to the native one is illustrated in Figure 15 (69), but its energy was 17 kcal/mol higher than the lowest-energy structure identified in the conformational search.

The distribution of energies obtained with each of the hydration models (69) is illustrated in Figure 16. It can be seen that the generated ensemble covers a wide range of RMSD's but a narrower range of energies around the lowest-energy value. This result represents one of the main difficulties in identifying the native-like structure. On the other hand, this means that the total energy, as a scoring function, must be extremely precise in order to distinguish the correct folded structure from wrong ones. This constitutes a challenge for improving existing force fields or for developing new potential functions. In addition, the small energy gap between basins containing quite different folds represents a challenge for search methods. We have subsequently applied the EDMC method to the villin headpiece (73) with results of similar quality. However, computations at pH 3.7 and 7.0 for the ten native-like structures satisfying the NMR-derived constraints indicate a substantial change in the charge distribution for each type of amino-acid residue with the change of pH. In particular, at pH 3.7 at which the NMR experiments were carried out, the lowest-energy conformation found during the conformational search satisfies ~70% of both the distance- and the dihedral-angle constraints, and possesses the characteristic packing of three phenylalanine residues that constitute the main part of the hydrophobic core of the molecule. The results of these computations indicate that an accurate description of the electrostatic interactions appears to be crucial for protein stability and consequently for an accurate prediction of the native state. It is not yet clear what the largest size protein is that can be treated by this all-atom EDMC procedure. Nevertheless, it is encouraging that an all-atom representation of the chain, and global optimization of the corresponding potential energy, can identify the native-like fold without resorting to knowledge-based information in the search procedure. Pending the results of ongoing computations with the EDMC procedure, we are currently using the hierarchical

Figure 15. Ribbon diagram of the superposition of the native fold and the conformation with the lowest C^a RMSD from the native fold, obtained with the SRFOPT solvation model. The energy of this conformation is 17 kcal/mol above that of the lowest-energy conformation.

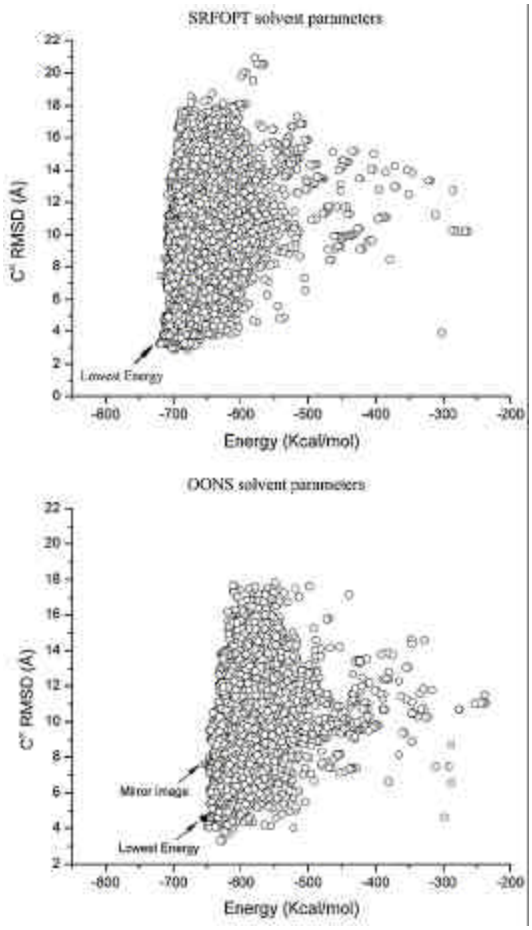

Figure 16. C^a RMSD vs the total energy, computed with the SRFOPT and OONS solvation models. The lowest energies belong to the lowest-energy structures shown in Figure 14. The mirror-image structure, computed with the OONS solvation model, was obtained as the lowest-energy structure in one of four runs, but its energy is higher by 1.4 kcal/mol than that of the lowest energy structure.

procedure introduced in Section 6 to treat proteins containing of the order of 100 residues with both a and ß folds.

5.5. Diffusion equation method (DEM)

A multiple-minima problem arises in the complex multi-dimensional potential energy surfaces, and various procedures, were developed (some of which are discussed above) for global optimization. A different approach has been taken by trying to smooth the energy surface and eliminate unwanted, higher-energy minima, leaving a descendant of the global minimum. As shown elsewhere (74), the DEM can accomplish this smoothing by solving the multi-dimensional diffusion equation

$$\Delta F = \partial F / \partial t \qquad (1)$$

to find the global minimum of a function f(**x**), with the boundary condition

$$F(\mathbf{x}, 0) = f(\mathbf{x}) \qquad (2)$$

where **x** is the vector of Cartesian coordinates, F is the deformed function, ? is the Laplacian operator, and t is time (really a deformation variable, rather than time, because this is not a time-dependent problem).

Figure 17 provides an example of a one-dimensional problem to illustrate the behavior of the DEM. This one-dimensional function contains two minima, and solution of the one-dimensional diffusion equation at various values of t eliminates the higher-energy, unwanted second minimum at some time t_0 = 0.25. However, the position of the minimum at t_0 = 0.25 has shifted from the position of the minimum at t = 0. Therefore, the global minimum at t = 0 is recovered by reversing the procedure. The minimum on the t_0 surface is not the minimum on the t = 0.15 surface; however, the former is close to the latter. Therefore, starting with the minimum on the t_0 surface, and minimizing the function on the t = 0.15 surface, the minimum on the t = 0.15 surface is obtained. By continuing the procedure of alternating solving the diffusion equation at lower t's with energy minimization, the function can be tracked back to the global minimum of the original function.

This procedure has worked on many complex mathematical functions (74) and on selected clusters of argon particles (75) that interact with a Leonard-Jones potential. The latter is one component of ECEPP. An example of three argon clusters is illustrated in Figure 18. The structure of the 55-particle cluster, with 159 degrees of freedom was found in ~400 sec on an IBM 3090 computer (75). A subsequent variant of the DEM, a self-consistent Basin-to-Deformed-Basin Mapping (SCBDBM) method (76, 77), using distance scaling (78), extended the search to a 100-particle argon cluster, requiring about 3.5 hours with 10 processors of an IBM SP2 supercomputer, and was applied successfully to all clusters smaller than 100 atoms.

The first application of the DEM to a peptide was made for the terminally-blocked alanine residue (79), erroneously called a "dipeptide". Its ECEPP (f, y) map is shown in Figure 19. The global minimum is at point C,

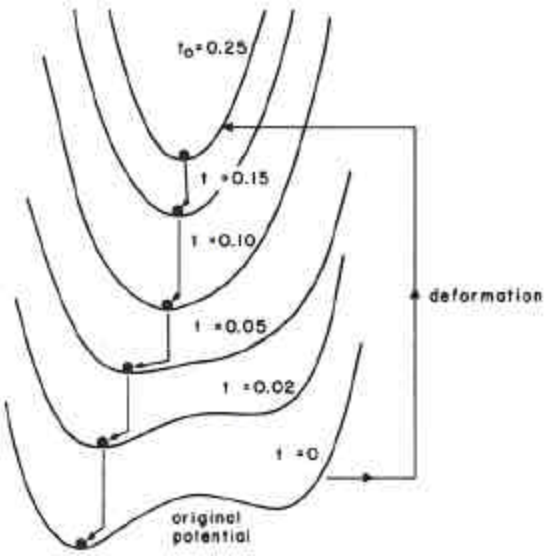

Figure 17. Illustration of the deformation of a one-dimensional potential function f(x), and of the reversing procedure. The deformation leads to the curve, at $t_o = 0.25$, with a unique minimum that is achievable from any point of the space by a simple minimization. Then, the reversing procedure (shown by the arrows directed downward) is applied by considering a sequence of the deformed curves at lower values of t. Each step of the procedure is followed by a minimization symbolized by a ball moving downhill from the minimum position of the upper curve and always reaching the position of the minimum in the lower curve. In the final step, the global minimum of f(x) is found.

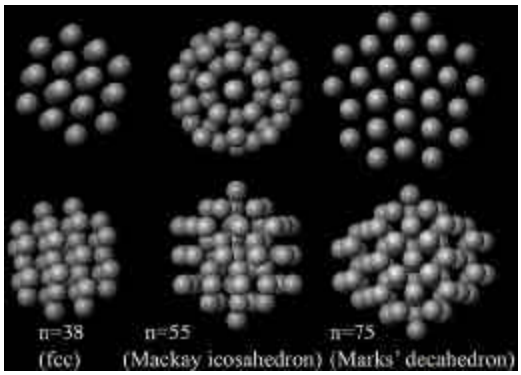

Figure 18. Minimum-energy geometries found by the DEM for argon clusters with n = 38, 55 and 75 atoms.

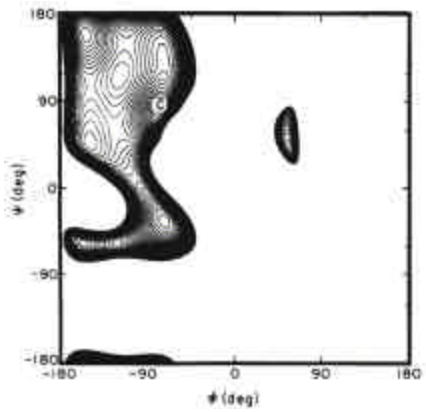

Figure 19. ECEPP (f, y) map of terminally-blocked alanine at t = 0. Point C is the position of the global minimum.

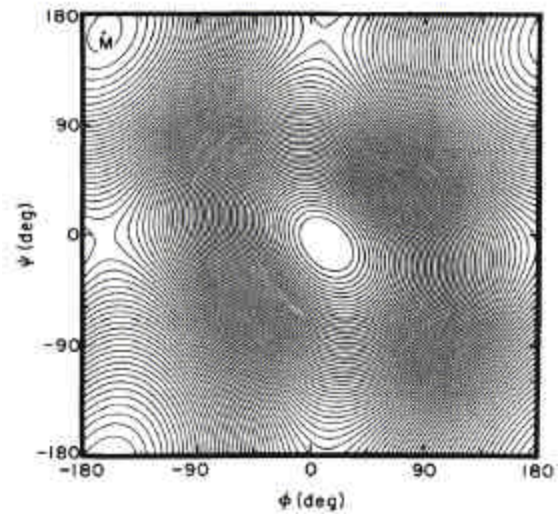

Figure 20. Deformed potential surface for terminally-blocked alanine at a particular time. The position of the unique minimum is indicated by M. There is a maximum at f ~ 15°, ? ~ -15°, and saddle points at f ~ 10°, ? ~ 165° and at f ~ -165°, ? ~ -10°.

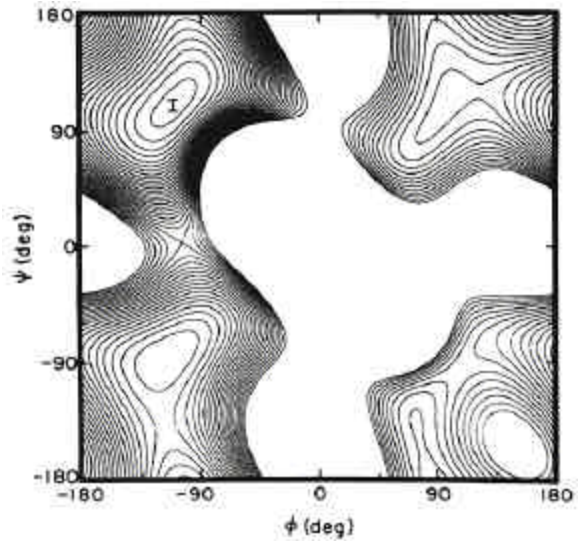

Figure 21. Deformed potential surface for terminally-blocked alanine at an intermediate time during the time reversal. The unique minimum of the higher-time map of Figure 20 has moved to the intermediate position I but (even though other minima may appear) there is no problem in distinguishing I from the position of the global minimum of the original ECEPP function, i.e. at t = 0. The contours in the lower right-hand corner correspond to a maximum.

which corresponds to the C_7^{eq} hydrogen-bonded structure; other higher-energy minima are also shown. Figure 20 illustrates an appropriately deformed surface after a suitable value of t (79). It contains only one minimum (at point M). The central region of the map is a maximum, and the other regions on the borders are saddle points. The reversal was carried out in 100 small steps at decreasing values of t. One of the intermediate maps is shown in Figure 21. The minimum at point M in Figure 20 has moved to the intermediate point I in Figure 21, but not yet to the global minimum at point C of Figure 19. Continuation of the reversing procedure to t = 0 led to the identical map of Figure 19, and the trajectory of movement of point M of Figure 20 to point C of Figure 19 is illustrated in Figure 22; the original map of Figure 19 was recovered in this reversal to t = 0.

Attempts to apply the DEM to longer peptides led to convergence problems which were overcome in the context of another global optimization problem, viz., the ab initio prediction of crystal structures, i.e., without using knowledge-based information such as the space group. The SCBDBM method (76, 77) was developed for this purpose; it essentially uses successive cycles of deformation and reversal. The SCBDBM method has also been applied to united-residue poly(L-alanine) chains with a length of up to 100 amino acid residues, and to locate low-energy conformations of the 10-55 fragment of the B-domain of staphyloccal protein A. An alternative approach, a Conformational Family Monte Carlo (CFMC) method (80), developed for global optimization of proteins, was also applied for calculations of crystal structures. The CFMC method maintains a database of low-energy conformations which are clustered into families. The conformations in this database are improved iteratively by a Metropolis-type Monte Carlo procedure together with energy minimization, in which the search is biased toward the regions of the lowest-energy families. By using the families instead of single conformations, CFMC coarse-grains the conformational space and exploits information about nearby low-energy states. CFMC has been applied to protein A and achieves the same performance as the CSA method discussed in section 6.2. The method is now used for the global optimization of crystal structures.

Deformation methods such as the DEM, however mathematically elegant, thus far appear to be numerically less efficient than stochastic methods such as MCM and EDMC. The deformation methods require more computational time than the stochastic methods to reach comparably low energies.

5.6. Global optimization of crystal structures

Motivated by the need to circumvent the problems that appeared in the DEM, the SCBDBM and CFMC methods were developed, and also applied to the prediction of crystal structures. This is another of many applications of global optimization in physics, and this section is a diversion to illustrate how our methodology can treat another problem of physical interest. It is especially challenging in light of statements in the literature (81, 82) that the global optimization problem has not yet been solved for predicting crystal structures, although these statements have been tempered by recent cautious optimism (83). As with our approach to protein-structure calculations, we use an ab initio approach, without using knowledge of the space group i.e., to predict both the arrangement of the molecules in the crystal and also the lattice constants.

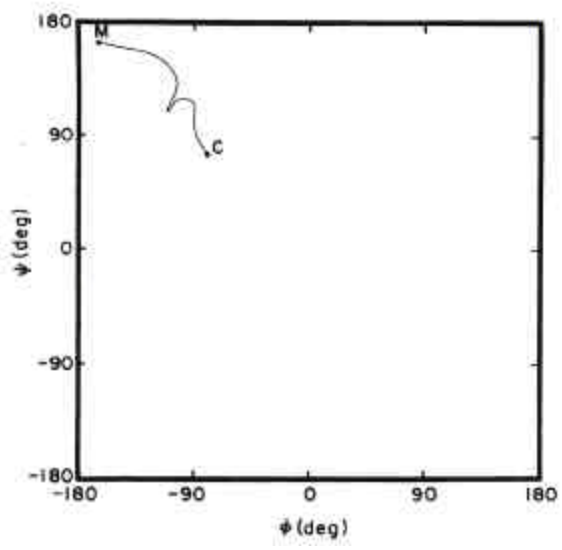

Figure 22. Trajectory of the global minimum for terminally-blocked alanine as t ? 0. Point M corresponds to the unique minimum in Figure 20, and point C corresponds to the global minimum of the original map (Figure 19) at t = 0.

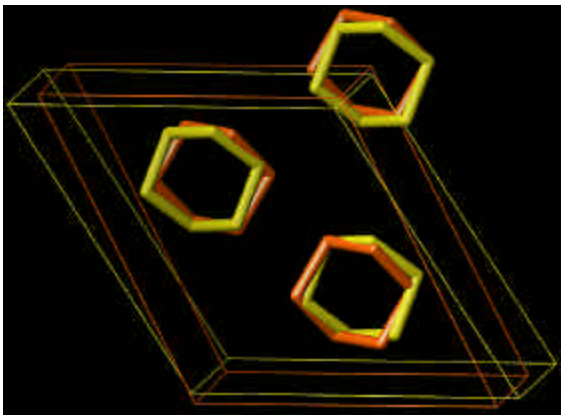

Figure 23. Computed structure for hexasulfur found by the DEM, superposed on the experimental structure.

The first application was made to a nonpolar crystal of S_6 molecules (84) in which only a Lennard-Jones potential is involved in the intermolecular interactions. The computed structure is shown in Figure 23, in which the coincidence of the computed and experimental molecular positions, as well as the lattice vectors, are reproduced quite well.

Application to a crystal of a benzene molecule with a partial-charge representation, but without a permanent dipole moment, required use of the Ewald summation to treat the electrostatic contribution (84), and led to the structure in Figure 24, where the familiar edge-to-face arrangement of the molecules is seen. Further calculations were carried out for crystals of polar molecules with permanent dipole moments (85), with results shown in Figure 25.

Participation in two recent blind tests on crystal structure prediction (86, 87), organized by the Cambridge Crystallographic Data Centre, provided a good test for our global search method and showed that it is efficient enough to predict crystals of rigid and flexible molecules if accurate potentials are used. When applied to crystals, the global search methods provide information about their potential energy surface, and therefore, can be used as a tool for evaluating potentials (85). The information they provide enabled us to develop a new global-optimization-based method for parameter optimization (88, 89).

While all the results of crystal calculations showed small deviations between experimental and calculated structures, global optimization of the potential energy can reduce these deviations by improving the parameters of the potential function to force it to lead to better agreement between computed and experimental structures. This global optimization-based approach is now being used to obtain an improved all-atom potential function (89-91).

6. HIERARCHICAL APPROACH TO PREDICT STRUCTURES OF LARGE PROTEIN MOLECULES

As pointed out in section 5.4, it is not yet clear whether the all-atom EDMC procedure can be applied, within a reasonable amount of computing time, to larger protein molecules containing of the order of 100-200 residues. Therefore, a hierarchical approach was developed to treat this problem (92-100), and hopefully overcome the suggestion that the ab initio prediction of protein structure may not be feasible in the foreseeable future (101-103).

In the hierarchical approach, global optimization is carried out by using a Conformational Space Annealing (CSA) method (104, 105) with a united-residue (UNRES) representation of the protein chain (94-96). This is the key stage of the hierarchical algorithm. It is designed to locate the *region* of the global minimum rapidly and efficiently. The lowest-energy structures obtained from the UNRES representation in this stage are then converted to the all-atom representation (106, 107), and a local search is carried out in the restricted region located with the UNRES/CSA approach. This is accomplished with the EDMC method and the ECEPP/3 force field (17), together with the SRFOPT hydration model (26). Initially, the backbone of the chain is constrained to the structure obtained by UNRES and CSA, but the constraints are gradually reduced as the calculations proceed.

6.1. The UNRES model

The UNRES model (94-96) is illustrated in Figure 26. It consists of a virtual-bond chain, i.e., a sequence of a-carbons (small empty circles) and united peptide groups (indicated by shaded circles) and united side chains (indicated by the shaded ellipsoids, whose size depends on the nature of the amino acid residue). The a-carbons are not centers of interaction, but merely serve to locate the backbone. The centers of interaction are the shaded circles and shaded ellipsoids, with a united-residue potential given by equation 3,

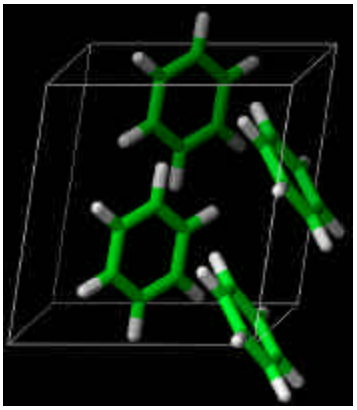

Figure 24. Computed structure for benzene found by the DEM.

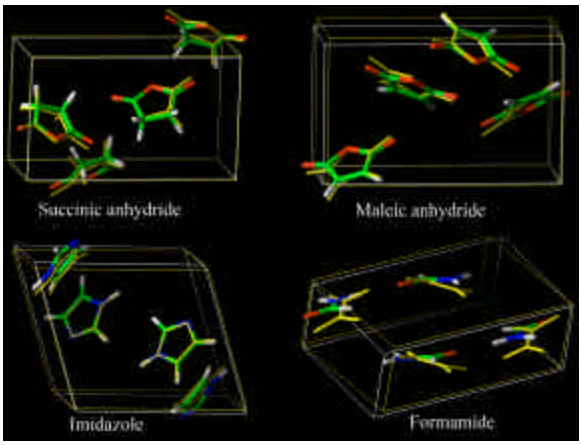

Figure 25. Computed structures for four compounds found by the SCBDBM method, superposed on the experimental structures.

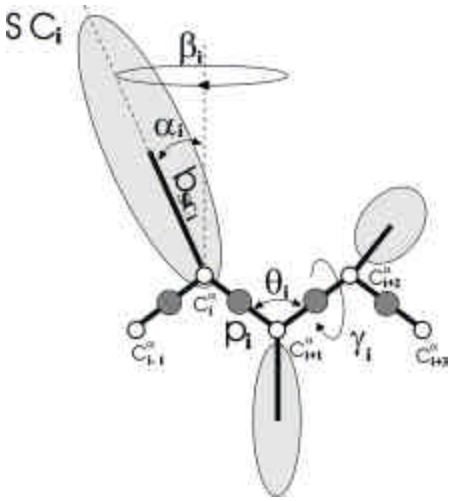

Figure 26. The UNRES model of polypeptide chains. The interaction sites are side-chain centroids of different sizes (SC) and the peptide-bond centers (*p*) indicated by shaded circles, whereas the a-carbon atoms (small empty circles) are introduced only to assist in defining the geometry. The virtual C^a-C^a bonds have a fixed length of 3.8 Å, corresponding to a trans peptide group; the virtual-bond (?) and dihedral (?) angles are variable. Each side chain is attached to the corresponding a-carbon with a fixed "bond length", b_{SC_i}, variable "bond angle", a_i, formed by SC_i and the bisector of the angle defined by C^a_{i-1}, C^a_i, and C^a_{i+1}, and with a variable "dihedral angle", ß$_i$, of counter-clockwise rotation about the C^a_{i-1}, C^a_i, C^a_{i+1} frame.

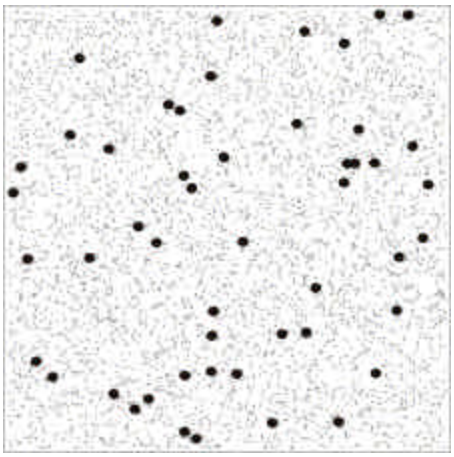

Figure 27. Schematic representation of the multi-dimensional space with all black and background points representing local minima in the CSA procedure.

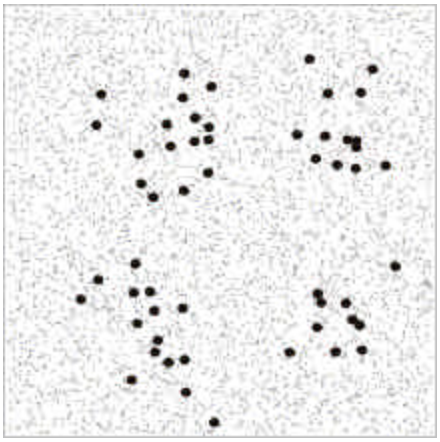

Figure 28. Schematic representation of intermediate stage of the CSA procedure, in which the black circles are coalescing here into four clusters.

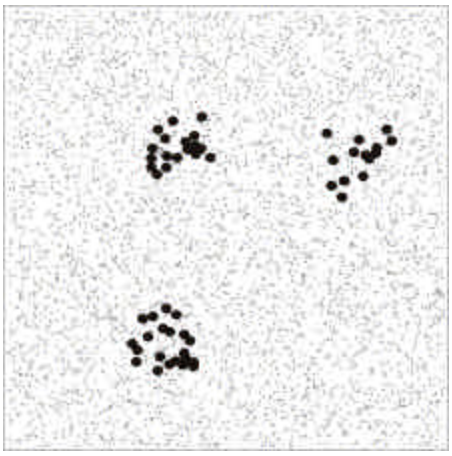

Figure 29. Schematic representation of the final stage of the CSA procedure.

$$U = \sum_{i<j} U_{sc(p)sc(p)_j} + w_{sc(p)} \sum_{i,j} U_{sc(p)j} + w_{el} \sum_{i<j-k} U_{el(p)_j} + w_{tor} \sum_i U_{tor}(\gamma_i) + w_b \sum_i U_b(\theta_i) + w_{rot} \sum_i U_{rot}(\alpha_{sc_i}, \beta_{sc_i}) + \sum_{m=2}^{n_{corr}} w_{corr}^{(m)} U_{corr}^{(m)} \quad (3)$$

where the successive terms represent side chain-side chain, side chain-peptide, peptide-peptide, torsional, bond-angle bending, side-chain angles a and ß, and multi-body (correlation) interactions, respectively. The w's are the relative weights of each term. The correlation terms arise from a cumulant expansion (97, 108) of the Restricted Free Energy (RFE) function (or potential of mean force) of the simplified chain obtained from the all-atom (e.g., ECEPP) energy surface by integrating out the secondary degrees of freedom. The variables to change conformation are the angles ($?_i$) between virtual bonds, the torsional angle ($?_i$) for rotation about the virtual bonds, and the position angle (a_i) and rotational angle ($ß_i$) of the side chains.

6.2. The CSA method

Figure 27 is a schematic representation of a multi-dimensional conformational space with its black circles and background points representing local energy minima. The CSA method (104, 105) is based essentially on a build-up and a genetic algorithm to force the 50-100 black circles to coalesce to the region of the global minimum. Figures 28 and 29 are schematic representations of this coalescence at various stages of the CSA procedure. Figure 29 represents a possible final stage of the procedure, illustrating three clusters of minima, with the global minimum presumably lying in one of these clusters.

Recently (109), we enhanced the power of the CSA method by introducing genetic operators that copy regular secondary-structure elements (e.g., ß-hairpins, a-helices, etc.) between conformations. It should be noted that these structural elements are those that are detected in the conformations found during the search and are not taken from structural databases. Additional enhancements include the treatment of disulfide bonds (110).

All UNRES minimum-energy conformations in the final clusters of Figure 29 are converted to the all-atom representation (106, 107), and the global optimization search is continued from these starting conformations with the EDMC procedure (68), as indicated above.

6.3. Initial tests of the UNRES/CSA procedure

Figures 30-32 illustrate the results of initial tests of the hierarchical procedure (111) on proteins of known structure. The search with protein A led to a structure (red) with a 3.8 Å RMSD from the experimental structure (blue). A mirror-image, higher-energy conformation (yellow) was also found. Figure 30 shows a scatter plot of C^{α} RMSD from the native structure vs. energy for intermediate conformations from the CSA simulation of protein A. It can be seen that more than one conformational family exists. A detailed analysis shows that the set can be divided into two families, one of which contains native-like structures and the other contains the mirror image of the experimental structure. With these initial results, the procedure was considered to be sufficiently robust for us to participate in the CASP3 (Critical Assessment of Protein Structure

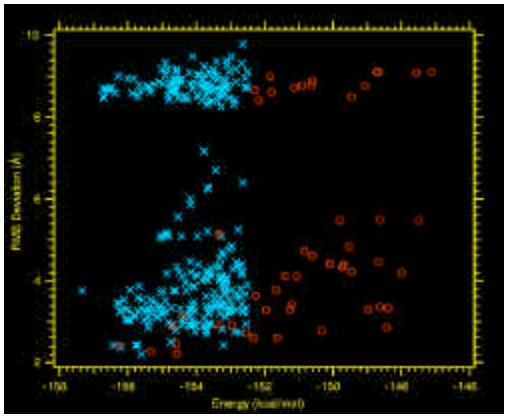

Figure 30. An intermediate stage of the CSA procedure applied to protein A. At a later stage, the red circles disappear, and two groups of minima represented by the blue crosses remain. The upper group is found to contain mirror images of the lower group.

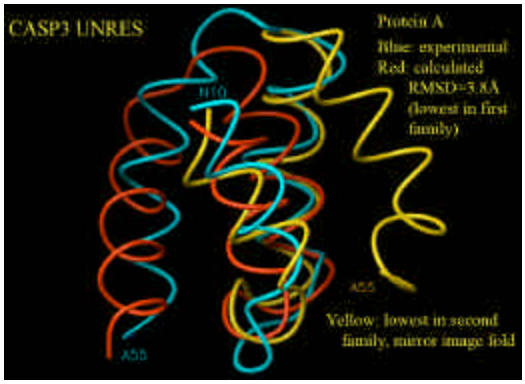

Figure 31. Superposition of the C^a traces of three structures of protein A. All residues of the blue and red structures were superposed, but only the first two helices of the blue and yellow structures were superposed. The yellow structure is the mirror image fold of the blue/red structures.

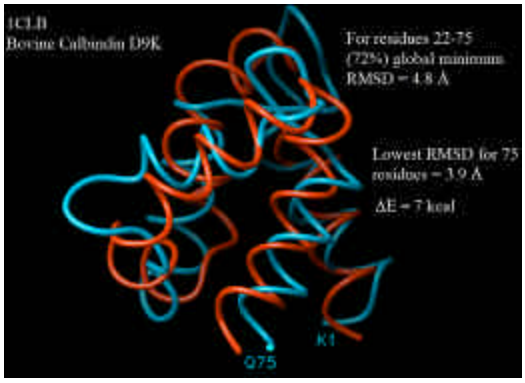

Figure 32. Superposition of the C^a trace of the calculated lowest-energy structure of apo calbindin D9K (red) on its NMR 1CLB structure (blue). The calculated structure is the mirror image of the native fold. Residues 22-75 were used for the superposition.

Prediction) blind prediction exercise in 1998. Subsections 6.4-6.7 illustrate the progressive improvement in UNRES and in the search procedure during our participation in successive CASP exercises.

6.4. CASP3 results

To provide a basis for interpreting the use of RMSD as a scoring function to evaluate the quality of a prediction, we cite the paper of Reva *et al.* (112) who asked the question, "What is the probability of a chance prediction of a protein structure at an RMSD of 6 Å?" They concluded that "the probability of obtaining a 6 Å RMSD by chance is so remote that, when such structures are obtained from a prediction algorithm, they should be considered successful."

In addition to the two initial tests of Figures 31 and 32, for proteins consisting of 46 and 75 residues, respectively, the results of two blind tests in CASP3 are illustrated for HDEA (Figure 33, 61 residues) and MarA (Figure 34, a 61-residue portion of the whole protein (113). The native structure of HDEA is a five-helix bundle with a long loop between the third (H-3) and the fourth (H-4) helix structures (Figure 33). Our computed model differs from the native structure by the packing of the 27-residue N-terminal portion. The N-terminal fragment, which contains helix H-2, is rotated by approximately 180°, resulting in an overall RMSD of 9.0Å. However, helices H-3 to H-5, a 61-residue portion, exhibits an RMSD of 4.2Å, and the superposed 27-residue N-terminal portion exhibits an RMSD of 2.9Å.

The CASP rules allow for the submission of five predicted structures (ranked by energy) for each target, i.e., amino acid sequence. The results of Figures 33 and 34 are two of 22 models submitted for seven targets. The CPU time with 64 IBM SP2 processors for this whole exercise was 1 hour for 37 amino acid residues to ~5 days for 140 amino acid residues. The computational problem is thus tractable with this amount of CPU time, pertaining to this older supercomputer, for proteins of this size range; newer supercomputers have greatly enhanced the efficiency of these computations.

It is of interest to point out that Mar A is a DNA-binding protein whose structure was determined from a complex of Mar A with a DNA chain (114). This information was not provided by the CASP3 organizers. Therefore, the calculations, carried out for the protein in the absence of the DNA, led to the collapse of the two domains which would bind to separate grooves of the DNA if the latter were present. However, a reasonable (6.0 Å) RMSD was obtained for one of the 61-residue domains, which is 53% of the whole protein.

In the evaluation of the ab initio results of CASP3, the judges (115) reported that "for protein HDEA (target 61)…the most impressive prediction was that of Scheraga's group using…*ab initio* methods…. Their method uses no information from sequence alignments, secondary structure prediction, or threading."

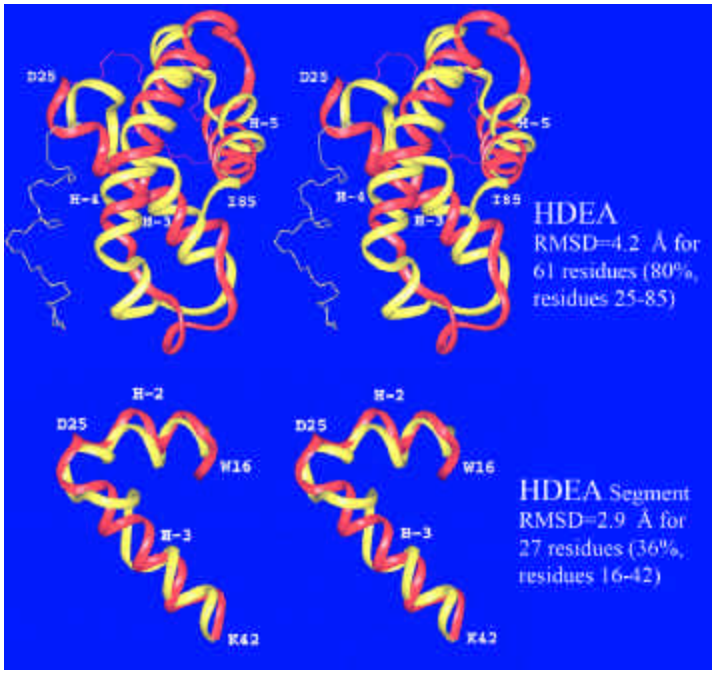

Figure 33. *Top*. Superposition of the crystal (red) and predicted (yellow) structures of HDEA. The C^{a} atoms of the fragment included between residues D25 and I85 were superposed. Helices 3, 4 and 5 are indicated as H-3, H-4 and H-5, respectively. *Bottom*. Superposition of the crystal (red) and predicted (yellow) structures of the 27-residue fragment (W16 to K42) of HDEA. Helices 2 and 3 are indicated as H-2 and H-3, respectively.

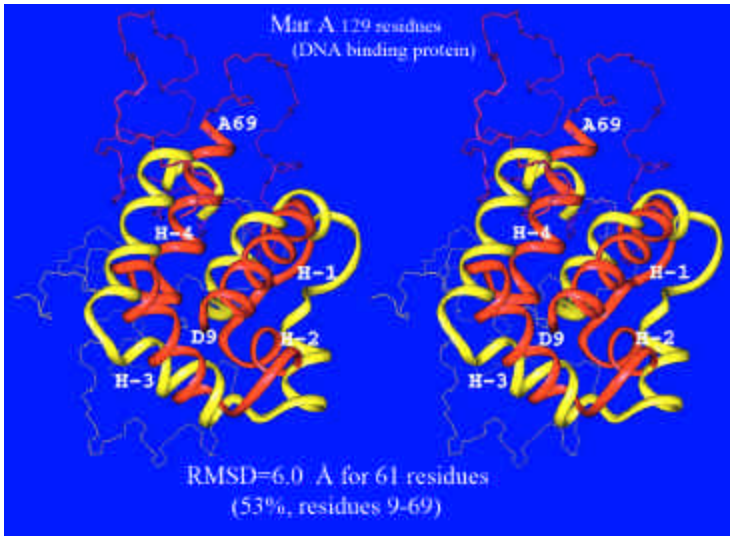

Figure 34. Superposition of the crystal (red) and predicted (yellow) structures of the Mar A N-terminal domain. The C^{a} atoms of residues D9 to A69 were superposed. Helices 1, 2, 3 and 4 are indicated as H-1, H-2, H-3 and H-4, respectively.

Figures 33 and 34 represent predictions for a-helical-type proteins. At the time of CASP3, our algorithm could not predict ß-structures, as illustrated on the left-hand side of the Figure 35 for betanova, whose structure was computed with the CASP3 UNRES force field. Therefore, to be able to predict ß as well as a structures, the UNRES force field was improved in preparation for CASP4 in 2000 (see subsection 6.5). The structure of betanova, computed with the improved (CASP4) UNRES force field is shown on the right-hand side of Figure 35, superimposed on the experimental (116) structure.

6.5. CASP4 results

The terms of the cumulant expansion have been evaluated analytically (97). Instead of showing complicated mathematical expressions for these terms, we represent them pictorially in Figure 36. The circles represent local interactions within a residue, and the vertical dashed lines

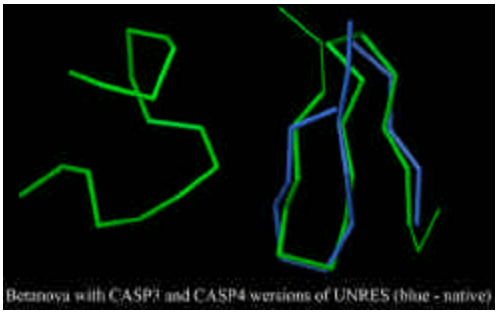

Figure 35. Betanova structures (green) computed with CASP3 (left) and CASP4 (right) versions of UNRES. The experimental structure is superposed on the computed structure (right).

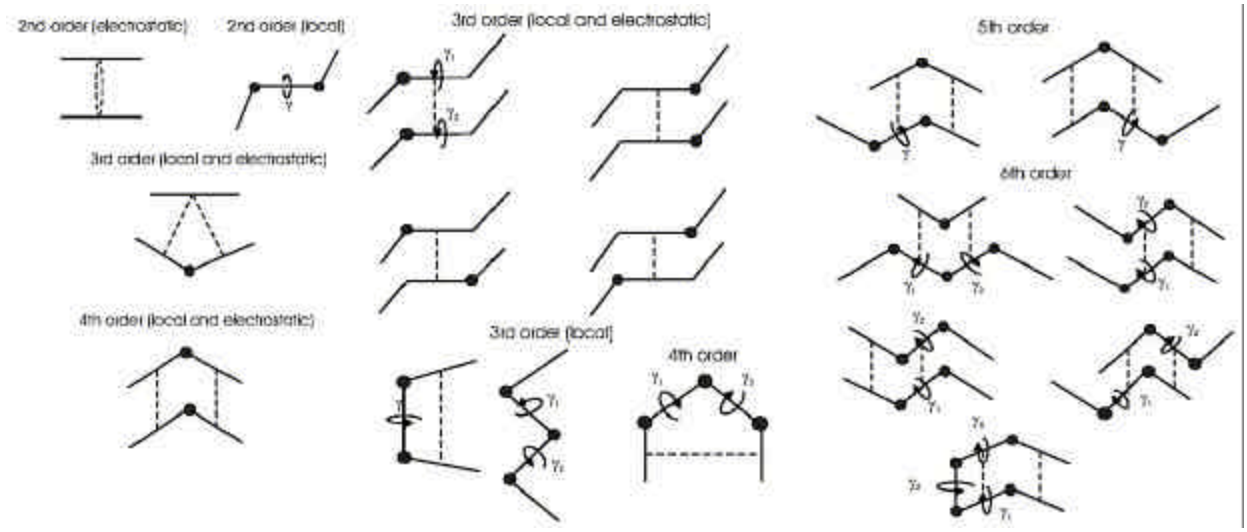

Figure 36. Graphical representations of the analytical forms of the terms of the cumulant expansion up to 6th order. Solid lines correspond to peptide groups involved in a correlation. Dashed lines correspond to backbone-electrostatic or hydrogen-bonding interactions, and circles correspond to backbone-local interactions.

represent electrostatic or hydrogen-bonding interactions between peptide groups. As seen, for example, in the 4th order term at the lower left-hand corner of Figure 36, it consists of four interacting (correlated) elements. If two such hydrogen bonds are formed, extra stability is provided, beyond the sum of two such (separate) hydrogen bonds. This analytical cooperative term was also obtained heuristically by Skolnick and coworkers (117).

In CASP3, only the terms of the left-hand column were used. However, in preparation for CASP4, all of the terms, up to 6th order, were used. It can be seen that the 6th order term in the lower right-hand corner of Figure 36 represents a feature that could lead to an anti-parallel two-stranded ß-structure. Likewise, the extended structure in the middle at the bottom of the middle column of Figure 36 should also facilitate the formation of ß-structure. With the Z-score re-optimization of the weights in equation 3, described in subsection 6.6, the results of Figure 37 were obtained for two combined tests on the a and ß structures. The improved UNRES force field then led to the results shown in Figure 38 for a test protein with an a and a ß segment.

The computed structure of the cyclic target 102 of CASP4 (118) is shown in Figure 39. The ends of the chain were found to be close to each other without imposing a loop-closing potential. The final structure of Figure 39 is a cyclic one, after a final energy minimization that did include a loop-closing potential. It is of interest to examine the RMSD for fragments of the structure of Figure 39. As can be seen in Figure 40, shorter fragments were obtained with a lower RMSD than the 4.2Å found for the complete 70-residue chain.

Figure 41 shows a superposition of a computed 68-residue fragment of target 98, with an RMSD of 5.9Å from the native structure. Figure 42 shows that the new CASP4 UNRES potential can now lead to a good structure for an all-ß protein with an RMSD of 6.5Å. The ß-sheets, as well as the connecting non-regular structured loop, are reproduced fairly well.

6.6. CASP5 results

To assess the possibility of a correlation between the torsions around successive virtual bonds, *ab initio*

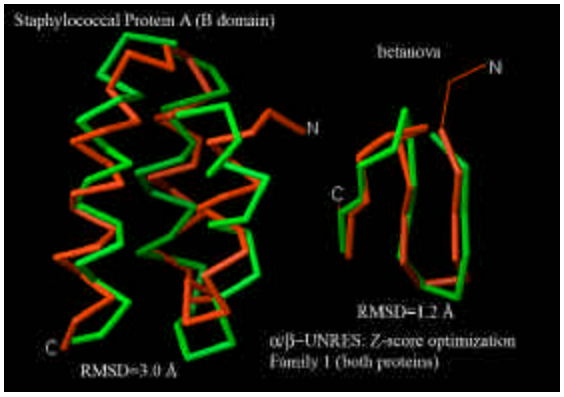

Figure 37. Results of Z-score optimization of the weights in equation 3 of the UNRES potential used in CASP4.

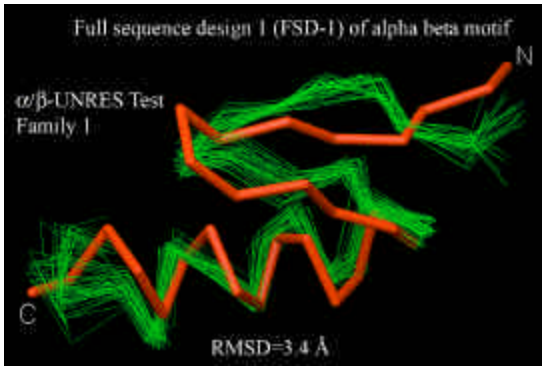

Figure 38. Test of the CASP4 force field on an a/ß protein. The computed structure (red) is superposed on a family of NMR structures (green).

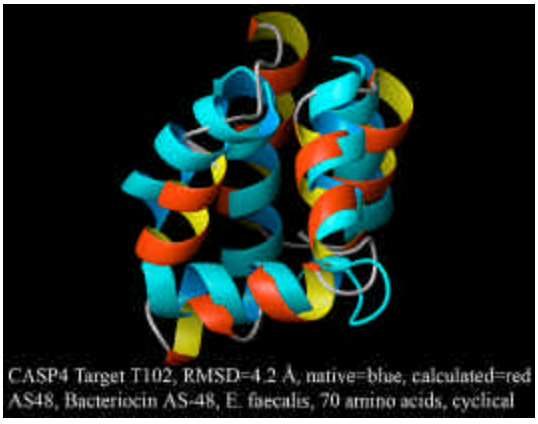

Figure 39. Computed structure (red) of CASP4 target 102 superposed on the experimental structure (blue).

quantum mechanical calculations were carried out on terminally-blocked glycine, alanine, and proline residues, and the respective torsional potentials were subsequently determined by numerical integration (29). It was found that there is an enhanced stabilization arising from such a correlation. Hence, a double-torsional term, $\sum_i U_{tord}(g_i, g_{i+1})$ shown in equation 4, was added to the UNRES potential.

$$U = \sum_{i} U_{SC(p)i} + w_{SCp} \sum_{i<j} U_{SC(p)i} + w_{ee} \sum_{i<j-1} U_{(p)i} + w_{ee} \sum_{i} U_{ee}(p_i) + w_{tord} \sum_{i} U_{tord}(\gamma_i, \gamma_{i+1}) + w_b \sum_{i} U_b(\theta_i) + w_{scl} \sum_{i} U_{scl}(\alpha_i, \beta_i) + \sum_{n=1}^{M} w_{ecm}^{n} U_{ecm}^{(n)} \quad (4)$$

The weights and also the other parameters of the energy terms of the cumulant expansion in equation 4 were re-computed by a gap (equation 5) and Z-score (equation 6) optimization.

$$\Delta E = \min_{i \in nat} E_i - \min_{i \in non-nat} E_i \quad (5)$$

$$Z = \frac{(1/N_{nat})\sum_{i=1}^{N_{nat}} E_i - (1/N_{non-nat})\sum_{i=1}^{N_{non-nat}} E_i}{\sqrt{(1/N_{non-nat})\sum_{i=1}^{N_{non-nat}} E_i^2 - \left[(1/N_{non-nat})\sum_{i=1}^{N_{non-nat}} E_i\right]^2}} \quad (6)$$

where E_i is the energy of the i[th] conformation, nat and non-nat denote the set of native-like and non-native structures, respectively, and N_{nat} and $N_{non-nat}$ are the numbers of conformations in the set of native-like and non-native structures, respectively.

Whereas previously the non-native structures were collected together in a group, it was found to lead to better results in the re-optimization of UNRES to separate the non-native structures into several levels as shown in Figure 43 (100). Level zero contains an ensemble of unstructured species, and level 1 contains an ensemble of species each of which contains one native element of secondary structure. Proceeding to the bottom level, each level contains successively more native regular structures, with the native packing of these regular structures increasing with movement toward the bottom level.

With this revised UNRES force field, the results shown in Figures 44 and 45 were obtained. Both targets shown in these Figures were classified as new folds. UNRES was able to predict a 70-residue segment of the C-terminal domain of target 149 [an (a + ß) protein]. Due to imperfections in the UNRES force field used in the CASP5 test, and also to speed up the search, we used secondary-structure information available from the public domain servers during the search. For target 129 (an a-helical protein), we were able to predict a 79-residue fragment, without using any knowledge-based information in the search.

6.7. Preparation for CASP6

In preparation for CASP6, the hierarchical Z-score approach of Figure 43 was modified and applied to a combined set of four proteins with representatives from different structural classes (119-121). The training proteins were 1E0G [(a + ß); 48 residues], 1E0L (ß; 28 residues), 1GAB (a; 47 residues) and 1IGD [(a + ß); 61 residues]. The force field was tested on a set of 66 proteins containing

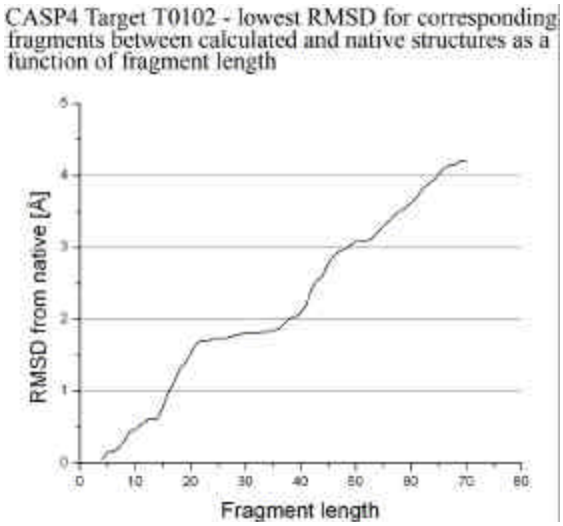

Figure 40. Evaluation of the performance on CASP4 target 102 as a function of fragment length.

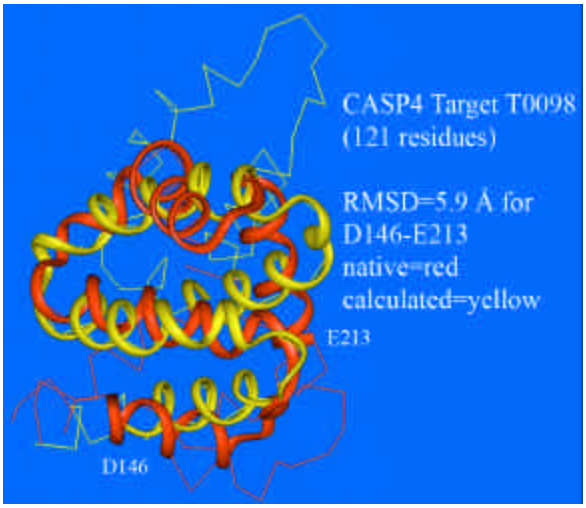

Figure 41. Superposition of calculated structure (yellow) on the native structure (red) for residues D146 to E213 of CASP4 target 98.

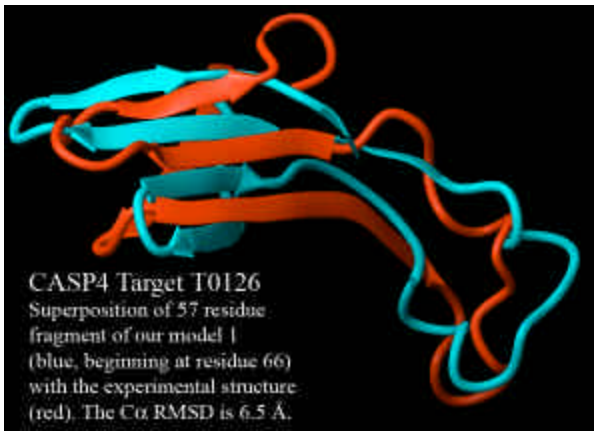

Figure 42. Superposition of calculated structure (blue) on the experimental structure (red) for a 57-residue fragment, beginning at residue 66.

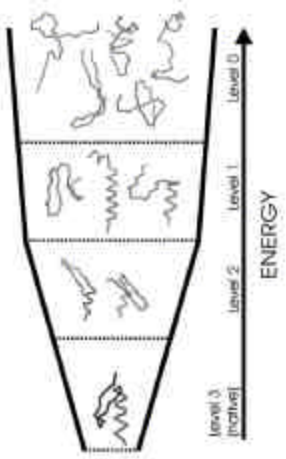

Figure 43. Schematic illustration of the energy levels of the 1FSD protein for optimizing the UNRES potential function. The energies of the conformations are required to decrease with their increasing "native likeness". The highest energy level (level 0) is occupied by structures with either no or non-native secondary structure. The next level (level 1) is occupied by the structures with one native secondary structure element (the N-terminal ß-hairpin or the C-terminal a-helix; the native-like structure fragments are indicated by thicker lines). Yet a lower energy level (level 2) has structures with both a-helix and ß-hairpin but no or incorrect packing of these two substructures and/or a shifted turn in the ß-hairpin. Finally, the native-like structures, with a-helix and ß-hairpin packed correctly, occupy the lowest energy level (level 3).

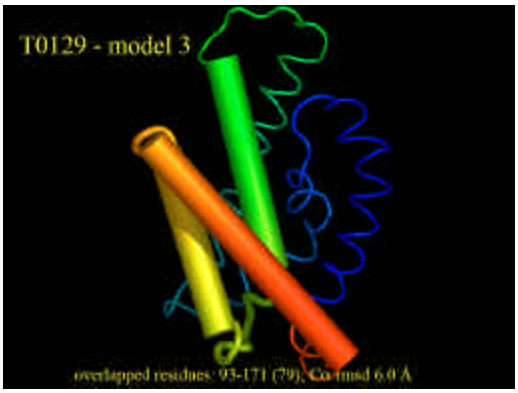

Figure 44. Prediction of the structure of residues 93-171 of CASP5 target 129.

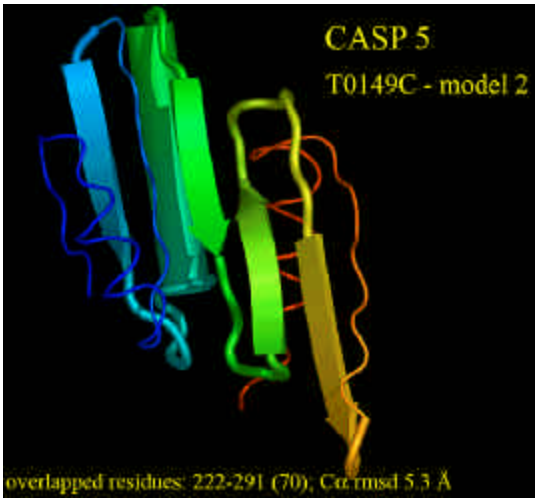

Figure 45. Prediction of the structure of residues 222-291 of CASP5 target 149.

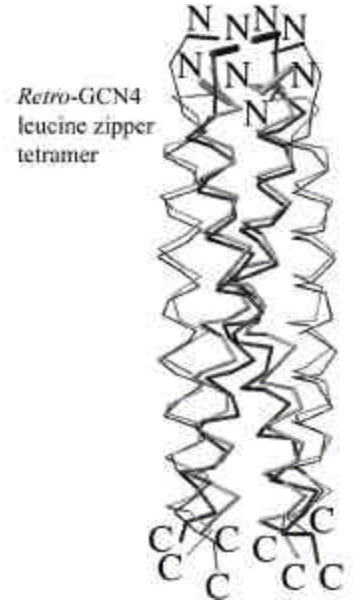

Figure 46. Lowest-energy structure of the rotationally symmetric retro-GCN4 leucine zipper (gray) superposed on the x-ray structure (black); C^a-coordinate RMSD = 2.34 Å.

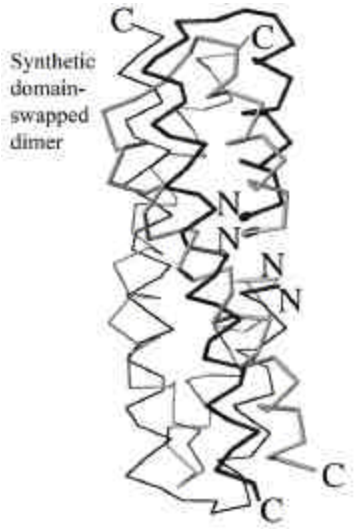

Figure 47. Lowest-energy structure calculated with rotational symmetry (gray) superposed on the experimental structure (black) for the synthetic domain-swapped dimer; C^a-coordinate RMSD = 5.65 Å.

28 to 147 residues and various structural types [26 a, 15 ß and 25 (a + ß)]. The average length of the segment predicted within 6 Å RMSD was 54 residues for a, 34 residues for ß, and 42 residues for (a + ß) proteins, which constituted 67%, 45%, and 55%, respectively, of the chain length. The longest predicted segments were 96 residues for a, 49 residues for ß, and 55 residues for (a + ß) proteins. These results are a considerable step forward compared to our earlier attempts to optimize the UNRES force field.

7. APPLICATION TO MULTIPLE CHAIN PROTEINS

The computations in section 6 were carried out for single-chain proteins. In order to apply the algorithm to multiple-chain proteins, the UNRES force field and CSA procedure were extended to treat proteins containing several chains (122, 123). Ideally, it should be possible to carry out the computations without knowing the number of chains in the protein or the symmetry relation between the chains. However, thus far, we have applied the procedure to proteins in which the number of chains is known, and they are related by rotational symmetry. Without this symmetry constraint, it has not yet been possible to achieve good results. However, work is in progress to try to remove the necessity of imposing a symmetry constraint.

With rotational symmetry imposed, the calculations were carried for the four-chain retro-GCN4 leucine zipper (122) shown in Figure 46 (with an RMSD of 2.34 Å) and for the two-chain synthetic domain-swapped dimer (122) shown in Figure 47 (with an RMSD of 5.65 Å). The calculated structures of the separated monomers of these two proteins (122), shown in Figures 48 and 49, respectively, do not resemble the structures of these monomers in the respective multiple-chain complexes.

8. CALCULATIONS OF FOLDING PATHWAYS

As pointed out in the Introduction (section 2), a second type of protein folding problem is the computation of the structural pathways by which the completely unfolded polypeptide chain proceeds to the folded native conformation. The problem is illustrated schematically in Figure 50 for an average folding pathway between a representative ensemble *i* of a completely unfolded protein (with no native contacts or native hydrogen bonds) and the final folded structure *f,* obtained by x-ray or NMR experiments. In this approach, structure *f* is not predicted, but has to be known in advance.

We have considered several theoretical approaches to determine folding pathways (10, 124-126). One of these makes use of the stochastic difference equation method of Elber *et al.* (126), and has been applied with a full-atom treatment to protein A which is shown in Figure 51. This method provides the sequence of formation of intermediate structures between *i* and *f* of Figure 50, but not the kinetics, there being no computation of the time scale for folding. This boundary value problem may be contrasted with an initial-value formulation in which one starts with the initial unfolded protein and uses molecular dynamics to compute the folding trajectory. The stochastic difference equation method avoids the problem that molecular dynamics requires femtosecond steps, and would consume an enormous amount of computer time to reach, say, a microsecond level, whereas most proteins (except for some very fast folders) fold in the millisecond-to-second time scale.

The method requires that the action S, as in equation 7 and 8,

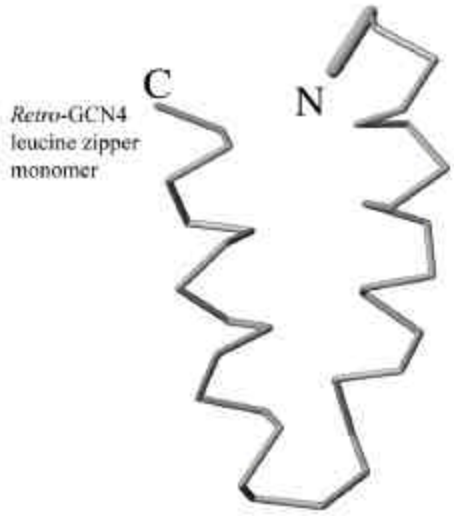

Figure 48. Lowest-energy calculated isolated monomer structure for the retro-GCN4 leucine zipper.

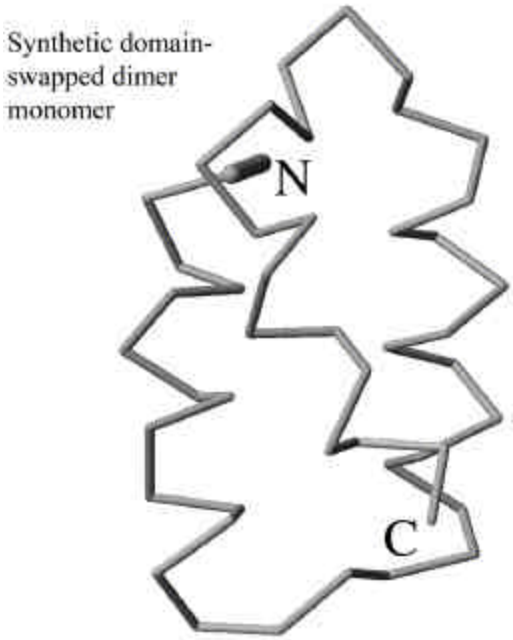

Figure 49. Lowest-energy calculated isolated monomer structure for the synthetic domain-swapped dimer.

Figure 50. Schematic representation of a folding pathway between given initial (*i*) and final (*f*) states. The *f* state is the known x-ray or NMR structure. The *i* state is an ensemble of unfolded polypeptide structures.

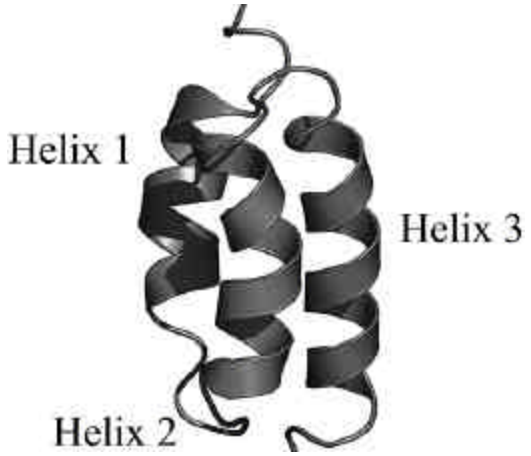

Figure 51. Native structure of protein A emphasizing the three helices whose folding order is computed in the pathway from *i* to *f* of Figure 50.

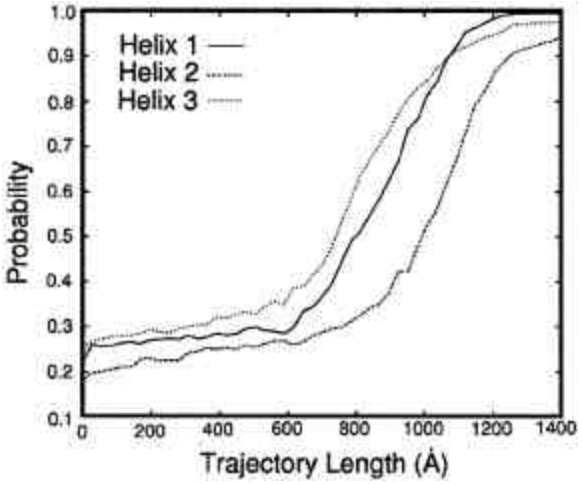

Figure 52. The fractions of amino acids in the helical conformation (probability) for each of the three helices of protein A. The fraction is followed as a function of the trajectory length measured in Ångstroms. The results are averaged over 130 folding trajectories. Helix 3 forms somewhat earlier than helices 1 and 2 in accordance with suggestive conclusions from experiments (127, 128). However, the difference between the rates of formation of helices 1 and 2 on the one hand, and helix 3 on the other, is not large.

$$S = \int_{Y_u}^{Y_f} \sqrt{2(E-U)}\, dl \qquad (7)$$

which is approximated by

$$S \cong \sum_i \sqrt{2(E-U)}\, \Delta l_{i,i+1} \qquad (8)$$

remain stationary along the length *l* of the pathway from *i* to *f*, where E is the total energy, *U* is the potential energy, and Y_u and Y_f are coordinates of the initial unfolded protein and of the final folded protein, respectively. To achieve a pathway with a stationary action, the function T of Equation 9 must be optimized.

$$T = \sum_i \left(\frac{\partial S / \partial Y_i}{\Delta l_{i,\, j+1}} \right)^2 \Delta l_{i,i+1} + l \sum_i \left(\Delta l_{i,i+1} < \Delta l > \right)^2 \qquad (9)$$

with

$$< \Delta l > = (1/N) \sum_i \Delta_{i,\, i+1} \qquad (10)$$

where the parameter ? is the strength of a penalty function that keeps all of the length elements, $\Delta l_{i,\, i+1}$, equal to the average length given by eq. 10, and equal to each other.

In treating protein A (10), an initial ensemble of 130 representatives of the unfolded protein were generated, and 130 separate trajectories to reach the final folded structure were computed, and an average was evaluated over all those trajectories. None of the 130 initial conformations had any native contacts or native hydrogen bonds.

The progress along the trajectories is shown in Figure 52, with separate curves for each of the three helices of protein A. It can be seen that helix 3 appears to fold first, followed by helix 1 and then by helix 2. To obtain a more detailed view of the folding pathways, the trajectory is divided into five equal segments, a-e, in Figure 53. It can be seen that, contrary to the current view that there is an initial hydrophobic collapse, followed by a slow rearrangement to form the native structure, panel a shows that there is a wide distribution of radii of gyration with very few native hydrogen bonds formed in the initial stage. Panel b, representing the second 20% of the trajectory, shows that only one or two additional (native) hydrogen bonds form but that there is still the wide distribution of radii of gyration. Only in panel c do we start to see a concomitant drop in the distribution of radii of gyration and a slight increase in the formation of native hydrogen bonds. This behavior continues into panels d and e in which the final radius of gyration and full complement of native hydrogen bonds appears. The same behavior is illustrated in Figure 54 in which it is seen that, in panel a, there are very few native contacts or native hydrogen bonds.

Future calculations of this type are planned for proteins for which experimental pathways are known. This will provide a direct experimental test of the theoretically calculated folding pathways.

9. CONCLUSIONS AND PERSPECTIVES

The results presented here, especially those of Figure 14, obtained from computations with an all-atom force field, demonstrate that the potential function and search procedure contain the essential ingredients to predict the folded structure from the amino acid sequence. Likewise, the results of Figures 53 and 54 suggest that the

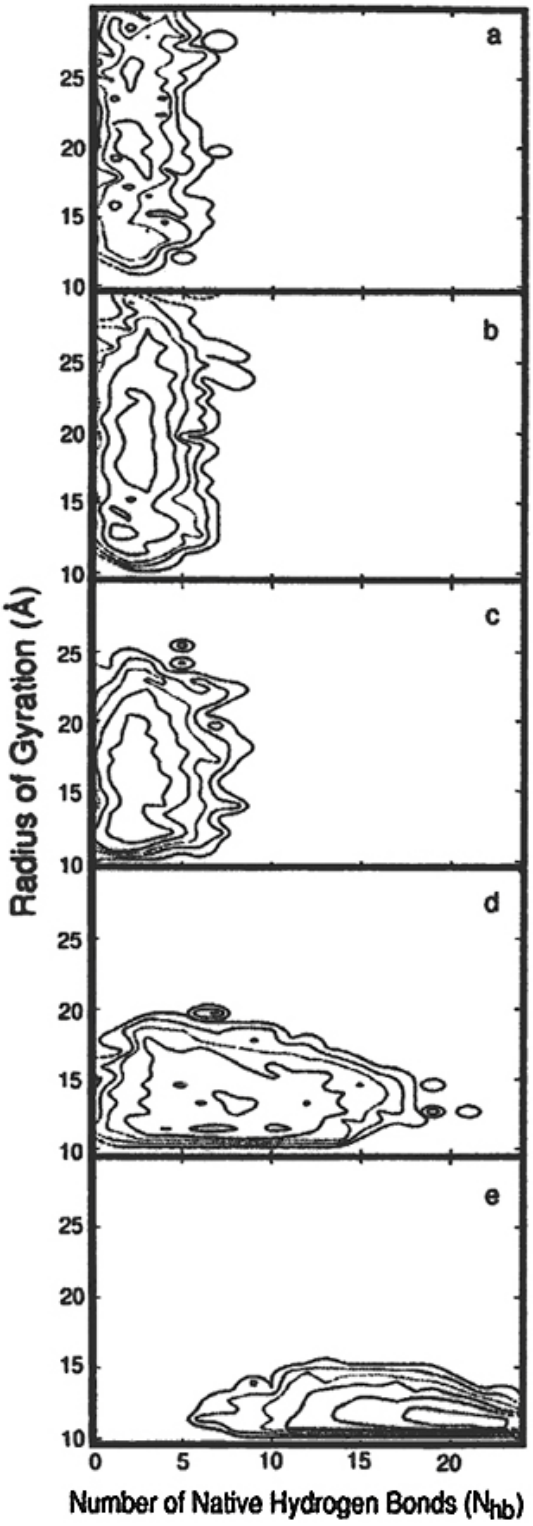

Figure 53. Two-dimensional "free energy profiles" are shown as a function of the radius of gyration and the number of hydrogen bonds at different sequential length slices (a-e) of the trajectory. The five snapshots in length are averaged over 130 trajectories and over the corresponding fifth of the trajectory length (e.g., the quantities in b are averaged over the second fifth). Sequential contour lines are separated by 1 kcal/mol

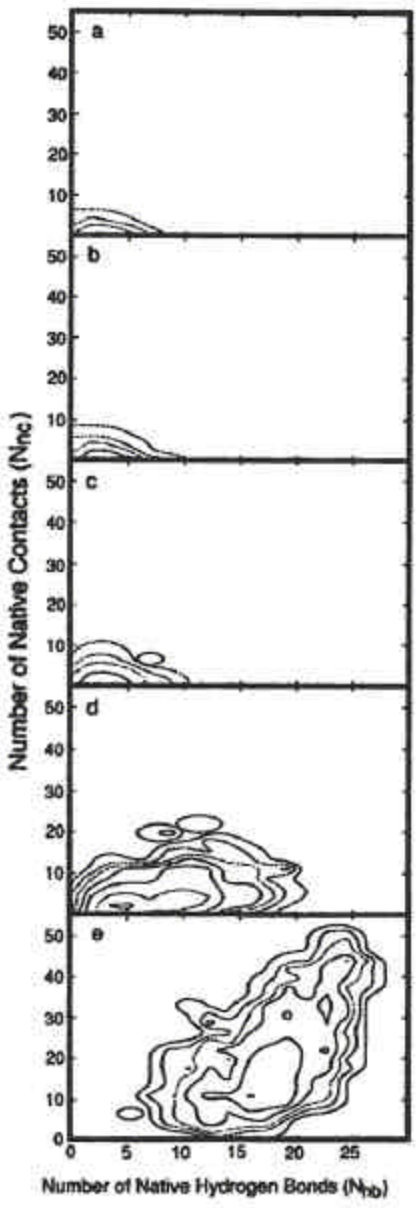

Figure 54. Two-dimensional "free energy profiles" are shown as a function of the number of hydrogen bonds and the number of native contacts for different length slices of the trajectory. The averages are over 130 trajectories and over the corresponding fifth of the trajectory length. Sequential contour lines are separated by 1 kcal/mol. The relatively slow progress of the folding process along two reaction coordinates in the early phases and the significant pick-up in speed in the last length segment should be noted.

available physics is sufficient to determine folding pathways. Presumably, refinements of the potential function and search procedures will lead to better agreement between theory and experiment. The foregoing statements apply so far only to proteins of the size of the 46-residue protein A.

At present, to extend the methodology to proteins in the 100-200 amino acid residue size range, we rely on the hierarchical approach with the UNRES and CSA search procedures, illustrated in Figures 26 and 27-29, respectively. Even this approach requires refinement of the UNRES potential and CSA search procedure. Such improvements are currently under investigation. It is not yet clear whether the all-atom force field will be applicable to proteins in the 100-200 residue size range.

10. ACKNOWLEDGMENTS

This work was supported by NIH (GM-14312) and NSF (MCB00-03722) grants. This research was conducted by using the resources of the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center.

11. REFERENCES

1. C. B. Anfinsen: Principles that govern the folding of protein chains. *Science* 181, 223-230 (1973)

2. T. E. Creighton and D. P. Goldenberg: Kinetic role of a meta-stable native-like two-disulfide species in the folding transition of bovine pancreatic trypsin inhibitor. *J Mol Biol* 179, 497-526 (1984)

3. J. S. Weissman and P. S. Kim: Reexamination of the folding of BPTI: Predominance of native intermediates. *Science* 253, 1386-1393 (1991)

4. M. Narayan, E. Welker, W. J. Wedemeyer and H. A. Scheraga: Oxidative folding of proteins. *Accts Chem Res* 33, 805-812 (2000)

5. H. A. Scheraga,: Structural studies of pancreatic ribonuclease. *Fed Proc* 26, 1380-1387 (1967)

6. H. A. Scheraga: Protein structure and function, from a colloidal to a molecular view (7th Linderstrøm Lang Lecture, Copenhagen, May 10, 1983). *Carlsberg Research Commun* 49, 1-55 (1984)

7. G. Némethy and H. A. Scheraga: Theoretical determination of sterically allowed conformations of a polypeptide chain by a computer method. *Biopolymers* 3, 155-184 (1965)

8. H. A. Scheraga: Calculations of conformations of polypeptides. *Adv Phys Org Chem* 6, 103-184 (1968)

9. H. A. Scheraga, J. Pillardy, A. Liwo, J. Lee, C. Czaplewski, D. R. Ripoll, W. J. Wedemeyer and Y.A. Arnautova: Evolution of physics-based methodology for exploring the conformational energy landscape of proteins. *J Comput Chem* 23, 28-34 (2002)

10. A. Ghosh, R. Elber and H. A. Scheraga: An atomically detailed study of the folding pathways of protein A with the stochastic difference equation. *Proc Nat Acad Sci USA* 99, 10394-10398 (2002)

11. N. Go and H. A. Scheraga: Analysis of the contribution of internal vibrations to the statistical weights of equilibrium conformations of macromolecules. *J Chem Phys* 51, 4751-4767 (1969)

12. N. Go and H. A. Scheraga: On the use of classical statistical mechanics in the treatment of polymer chain conformation. *Macromolecules* 9, 535-542 (1976)

13. H. A. Scheraga: Predicting three-dimensional structures of oligopeptides. In: Reviews in Computational Chemistry, Vol. 3, Eds: K. B. Lipkowitz and D. B. Boyd, VCH Publ, New York, 73-142 (1992)

14. F. A. Momany, R. F. McGuire, A. W. Burgess and H. A. Scheraga: Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. *J Phys Chem* 79, 2361-2381 (1975)

15. M. J. Sippl, G. Némethy and H. A. Scheraga: Intermolecular potentials from crystal data. 6. Determination of empirical potentials for O-H···O=C hydrogen bonds from packing configurations. *J Phys Chem* 88, 6231-6233 (1984)

16. G. Némethy, M. S. Pottle and H. A. Scheraga: Energy parameters in polypeptides. 9. Updating of geometrical parameters, nonbonded interactions, and hydrogen bond interactions for the naturally occurring amino acids. *J Phys Chem* 87, 1883-1887 (1983)

17. G. Némethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey and H. A. Scheraga: Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J Phys Chem* 96, 6472-6484 (1992)

18. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman: A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc* 117, 5179-5197 (1995)

19. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus: CHARMm: A program for macromolecular energy, minimization, and dynamics calculations. *J Comp Chem* 4, 187-217 (1983)

20. M.-J. Hwang, X. Ni, M. Waldman, C. S. Ewig and A. T. Hagler: Derivation of Class II force fields. VI. Carbohydrate compounds and anomeric effects. *Biopolymers* 45, 435-468 (1998)

21. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein: Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79, 926-935 (1983)

22. W. L. Jorgensen and J. D. Madura: Temperature and size dependence for Monte Carlo simulations of TIP4P water. *Mol Phys* 56, 1381-1392 (1985)

23. K. D. Gibson and H. A. Scheraga: Minimization of polypeptide energy. I. Preliminary structures of bovine pancreatic ribonuclease S-peptide. *Proc Natl Acad Sci US* 58, 420-427 (1967)

24. T. M. Ooi, Oobatake, G. Némethy and H. A. Scheraga: Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc Natl Acad Sci USA* 84, 3086-3090 (1987). *Erratum: ibid* 84, 6015 (1987)

25. Y. K. Kang, K. D. Gibson, G. Némethy and H. A. Scheraga: Free energies of hydration of solute molecules. 4. Revised treatment of the hydration shell model. *J Phys Chem* 92, 4739-4742 (1988)

26. J. Vila, R. L. Williams, M. Vasquez and H. A. Scheraga: Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitor, *Proteins: Structure. Function and Genetics* 10, 199-218 (1991)

27. D. Eisenberg and A. D. McLachlan: Solvation energy in protein folding and binding. *Nature* 319, 199-203 (1986)

28. G. D. Hawkins, C. J. Cramer and D. G. Truhlar: Pairwise solute descreening of solute charges from a dielectric medium. *Chem Phys Lett* 246, 122-129 (1995)

29. S. Oldziej, U. Kozlowska, A. Liwo and H. A. Scheraga: Determination of the potentials of mean force for rotation about Ca-Ca virtual bonds in polypeptides from the ab initio energy surfaces of terminally-blocked glycine, alanine, and proline. *J Phys Chem A* 107, 8035-8046 (2003)

30. K. Maksimiak, S. Rodziewicz-Motowidlo, C. Czaplewski, A. Liwo and H. A. Scheraga: Molecular simulation study of the potentials of mean force for the interactions between models of like-charged and between charged and nonpolar amino acid side chains in water. *J Phys Chem B* 107, 13496-13504 (2003)

31. A. Liwo, S. Oldziej, C. Caplewski, U. Kozlowska and H. A. Scheraga: Parameterization of backbone-electrostatic and multibody contributions to the UNRES force field for protein-structure prediction from ab initio energy surfaces of model systems. *J Phys. Chem B* 108, 9421-9438 (2004)

32. I. K. Roterman, M. H. Lambert, K. D. Gibson and H. A. Scheraga: A comparison of the CHARMM, AMBER and ECEPP potentials for peptides. II. f, ? maps for N-acetyl alanine N'-methyl amide: comparisons, contrasts and simple experimental tests. *J Biomolec Structure & Dynamics* 7, 421-453 (1989)

33. H. A. Scheraga, J. Lee, J. Pillardy, Y.-J. Ye, A. Liwo, and D.R. Ripoll: Surmounting the multiple-minima

problem in protein folding. *J Global Optimization* 15, 235-260 (1999)

34. I. Simon, G. Némethy and H. A. Scheraga: Conformational energy calculations of the effects of sequence variations on the conformations of two tetrapeptides. *Macromolecules* 11, 797-804 (1978)

35. M. R Pincus, R. D. Klausner and H. A. Scheraga: Calculation of the three-dimensional structure of the membrane-bound portion of melittin from its amino acid sequence. *Proc Natl Acad Sci USA* 79, 5107-5110 (1982)

36. M. Vasquez and H. A. Scheraga: Use of build-up and energy-minimization procedures to compute low-energy structures of the backbone of enkephalin. *Biopolymers* 24, 1437-1447 (1985)

37. K. D. Gibson and H. A. Scheraga: Revised algorithms for the build-up procedure for predicting protein conformations by energy minimization. *J Comput Chem* 8, 826-834 (1987)

38. G. N. Ramachandran, C. Ramakrishnan and V. Sasisekharan: Stereochemistry of polypeptide chain configurations. *J Mol Biol* 7, 95-99 (1963)

39. M. Dygert, N. Go and H. A. Scheraga: Use of a symmetry condition to compute the conformation of gramicidin S. *Macromolecules* 8, 750-761 (1975)

40. N. Go and H. A. Scheraga: Ring closure in chain molecules with Cn, I or S2n symmetry. *Macromolecules* 6, 273-281 (1973)

41. G. M. J. Schmidt, D. C. Hodgkin and B. M. Oughton: A crystallographic study of some derivatives of gramicidin S. *Biochem. J* 65, 744-750 (1957)

42. P. A. Mirau and F. A. Bovey: 2D and 3D NMR studies of polypeptide structure and function, Abstracts, 199th ACS meeting, Polymer Division. *Boston* 206 (1990)

43. M. H. Miller and H. A. Scheraga: Calculation of the structures of collagen models. Role of interchain interactions in determining the triple-helical coiled-coil conformation. I. Poly(glycyl-prolyl-prolyl). *J Polymer Sci Polymer Symposia No* 54 171-200 (1976)

44. G. N. Ramachandran and G. Kartha: Structure of collagen. *Nature* 176, 593-595 (1955)

45. A. Rich and F. H. C. Crick: The molecular structure of collagen. *J Mol Biol* 3, 483-506 (1961)

46. A. Yonath and W. Traub: Polymers of tripeptides as collagen models. IV. Structure analysis of poly(L-prolyl-glycyl-L-proline). *J Mol. Biol* 43, 461-477 (1969)

47. K. Okuyama, N. Tanaka, T. Ashida, M. Kakudo, S. Sakakibara and Y. Kishida: An x-ray study of the synthetic polypeptide (Pro-Pro-Gly)10. *J Mol. Biol* 72, 571-576 (1972)

48. R. Z. Kramer, L. Vitagliano, J. Bella, R. Berisio, L. Mazzarella, B. Brodsky, A. Zagari and H. M. Berman: X-ray crystallographic determination of a collagen-like peptide with the repeating sequence (Pro-Pro-Gly). *J Mol Biol* 280, 623-638 (1998)

49. R. Berisio, L. Vitagliano, L. Mazzarella and A. Zagari: Crystal Structure of the collagen triple helix model [(Pro-Pro-Gly)10]3. *Protein Science* 11, 262-270 (2002)

50. G. Némethy, M. H. Miller and H. A. Scheraga: Calculation of the structures of collagen models. Role of interchain interactions in determining the triple-helical coiled-coil conformation. 4. Poly(glycyl-alanyl-prolyl). *Macromolecules* 13, 914-919 (1980)

51. M. H. Miller, G. Némethy and H. A. Scheraga: Calculation of the structures of collagen models. Role of interchain interactions in determining the triple-helical coiled-coil conformation. 2. Poly(glycyl-prolyl-hydroxyprolyl). *Macromolecules* 13, 470-478 (1980)

52. M. H. Miller, G. Némethy and H. A. Scheraga: Calculation of the structures of collagen models. Role of interchain interactions in determining the triple-helical coiled-coil conformation. 3. Poly(glycyl-prolyl-alanyl). *Macromolecules* 13, 910-913 (1980)

53. L. Vitagliano, G. Némethy, A. Zagari and H. A. Scheraga: Stabilization of the triple-helical structure of natural collagen by side-chain interactions. *Biochemistry* 32, 7354-7359 (1993)

54. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller: Equation-of-state calculations by fast computing machines. *J Chem Phys* 21, 1087-1092 (1953)

55. Z. Li and H. A. Scheraga: Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc Natl Acad Sci USA* 84, 6611-6615 (1987)

56. Z. Li and H. A. Scheraga: Structure and free energy of complex thermodynamic systems. *J Molec Structure (Theochem)* 179, 333-352 (1988)

57. T. L. Blundell, L., Hearn, I. J. Tickle, R. A. Palmer, B. A. Morgan, G. D. Smith and J. F. Griffin: Crystal structure of [Leu5]enkephalin. *Science* 205, 220 (1979)

58. I. L. Karle, J. Karle, D. Mastropaolo, A. Camerman and N. Camerman: [Leu5]enkephalin: four cocrystallizing conformers with extended backbones that form an antiparallel ß-sheet. *Acta Cryst B* 39, 625-637 (1983)

59. L. Glasser and H. A. Scheraga: Calculations on crystal packing of a flexible molecule, Leu-enkephalin. *J Mol Biol* 199, 513-524 (1988)

60. E. O. Purisima and H. A. Scheraga: An approach to the multiple-minima problem by relaxing dimensionality. *Proc Natl Acad Sci USA* 83, 2782-2786 (1986)

61. D. R Ripoll and H. A. Scheraga: The multiple-minima problem in the conformational analysis of polypeptides. III. An electrostatically driven Monte Carlo method; tests on enkephalin. *J Protein Chem* 8, 263-287 (1989)

62. K. A. Olszewski, L. Piela and H. A. Scheraga: Mean-field theory as a tool for intramolecular conformational optimization. 1. Tests on terminally-blocked alanine and Met-enkephalin. *J Phys Chem* 96, 4672-4676 (1992)

63. J. Lee and H.A. Scheraga: Conformational space annealing by parallel computations: Extensive conformational search of Met-enkephalin and of the 20-residue membrane-bound portion of melittin. *Intl J of Quantum Chem* 75, 255-265 (1999)

64. L. Piela and H. A. Scheraga: On the multiple-minima problem in the conformational analysis of polypeptides. I. Backbone degrees of freedom for a perturbed a-helix. *Biopolymers* 26, S33-S58 (1987)

65. D. R. Ripoll, J. A. Vila and H. A. Scheraga: How well aligned are the backbone dipoles with the electrostatic field in native folds? In preparation (2004)

66. D. R. Ripoll and H. A. Scheraga: On the multiple-minima problem in the conformational analysis of polypeptides. II. An electrostatically driven Monte Carlo method-tests on poly(L-alanine), *Biopolymers* 27, 1283-1303 (1988)

67. D. R. Ripoll and H. A. Scheraga: On the multiple-minima problem in the conformational analysis of polypeptides. IV. Application of the electrostatically driven Monte Carlo method to the 20-residue membrane-bound portion of melittin. *Biopolymers* 30, 165-176 (1990)

68. D. R. Ripoll, A. Liwo, and H. A. Scheraga: New developments of the electrostatically driven Monte Carlo method: Test on the membrane-bound portion of melittin. *Biopolymers* 46, 117-126 (1998)

69. J. A. Vila, D. R. Ripoll and H. A. Scheraga: Atomically detailed folding simulation of the B domain of staphylococcal protein A from random structures. *Proc Natl Acad Sci USA* 100, 14812-14816 (2003)

70. J. Lee, J. Pillardy, C. Czaplewski, Y. Arnautova, D. R. Ripoll, A. Liwo, K. D. Gibson, R. J. Wawak and H. A. Scheraga: Efficient parallel algorithms in global optimization of potential-energy functions. *Comput Physics Commun* 128, 399-411 (2000)

71. Y. Duan and P. A. Kollman: Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* 282, 740-744 (1998)

72. B. Zagrovic, C. D. Snow, M. R. Shirts and V. J. Pande: Simulation of folding for a small a-helical protein in atomistic detail using worldwide-distributed computing. *J Mol Biol* 323, 927-937 (2002)

73. D. R. Ripoll, J. A. Vila and H. A. Scheraga: Folding of the villin headpiece subdomain from random structures. Analysis of the charge distribution as a function of pH. *J Mol Biol* 339, 915-925 (2004)

74. L. Piela, J. Kostrowicki and H. A. Scheraga: The multiple-minima problem in the conformational analysis of molecules. Deformation of the potential energy hypersurface by the diffusion equation method. *J Phys Chem* 93, 3339-3346 (1989)

75. J. Kostrowicki, L. Piela, B. J. Cherayil and H. A. Scheraga: Performance of the diffusion equation method in searches for optimum structures of clusters of Lennard-Jones atoms. *J Phys Chem* 95, 4113-4119 (1991)

76. J. Pillardy, A. Liwo, M. Groth and H. A. Scheraga - An efficient deformation-based global optimization method for off-lattice polymer chains; self-consistent basin-to-deformed-basin mapping (SCBDBM). Application to united-residue polypeptide chains. *J Phys Chem* B103, 7353-7366 (1999)

77. J. Pillardy, A. Liwo, and H. A. Scheraga - An efficient deformation-based global optimization method [self-consistent basin-to-deformed-basin mapping (SCBDBM)]. Application to Lennard-Jones atomic clusters. *J Phys Chem* A103, 9370-9377 (1999)

78. J. Pillardy and L. Piela: Smoothing techniques of global optimization. The distance-scaling method in searches for the most stable Lennard-Jones atomic clusters. *J Comp Chem* 18, 2040-2049 (1997)

79. J. Kostrowicki and H. A. Scheraga: Application of the diffusion equation method for global optimization to oligopeptides. *J Phys Chem* 96, 7442-7449 (1992)

80. J. Pillardy, C. Czaplewski, W.J. Wedemeyer and H.A. Scheraga: Conformation-Family Monte Carlo (CFMC): An efficient computational method for identifying the low-energy states of a macromolecule. *Helv Chim Acta* 83, 2214-2230 (2000)

81. J. Maddox: Crystals from first principles. *Nature* 335, 201 (1988)

82. A. Gavezzotti: Are crystal structures predictable? *Acc Chem Res* 27, 309-314 (1994)

83. J. D. Dunitz: Are crystal structures predictable? *Chem Comm* 545-548 (2003)

84. R. J. Wawak, J. Pillardy, A. Liwo, K. D. Gibson and H. A. Scheraga: Diffusion equation and distance scaling methods of global optimization: Applications to crystal structure prediction. *J Phys Chem* 102, 2904-2918 (1998)

85. J. Pillardy, R. J. Wawak, Y. A. Arnautova, C. Czaplewski and H. A. Scheraga: Crystal structure prediction by global optimization as a tool for evaluating potentials: Role of the dipole moment correction term in successful predictions. *J Am Chem Soc* 122, 907-921 (2000)

86. W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. A. Dzyabchenko, P. Erk, A. Gavezzotti, D. W. M. Hofmann, F J. J. Leusen, J. P. M. Lommerse, W. T. M. Mooij, S. L. Price, H. Scheraga, B. Schweizer, M. V. Schmidt, B. P. van Eijck, P. Verwer and D. E. Williams: Crystal structure prediction of small organic molecules: a second blind test. *Acta Crystallogr B* 58, 647-661 (2002)

87. W. D. S. Motherwell *et al*: Acta Cryst. B. In preparation (2004)

88. J. Pillardy, Y.A. Arnautova, C. Czaplewski, K. D. Gibson, and H. A. Scheraga: Conformation-family Monte Carlo: A new method for crystal structure prediction. *Proc Natl Acad Sci USA* 98, 12351-12356 (2001)

89. Y. A. Arnautova, J. Pillardy, C. Czaplewski and H. A. Scheraga: Global optimization-based method for deriving intermolecular potential parameters for crystals. *J Phys Chem B* 107, 712-723 (2003)

90. Y. A. Arnautova, A. Jagielska, J. Pillardy and H. A. Scheraga: Derivation of a new force field for crystal-structure prediction using global optimization: nonbonded potential parameters for hydrocarbons and alcohols. *J Phys Chem B* 107, 7143-7154 (2003)

91. A. Jagielska, Y. A. Arnautova and H. A. Scheraga: Derivation of a new force field for crystal,: structure prediction using global optimization: nonbonded potential parameters for amines, imidazoles, amides and carboxylic acids. *J Phys Chem B* 108, 12181-12196 (2004)

92. A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky and H. A. Scheraga: Calculation of protein backbone geometry from a-carbon coordinates based on peptide-group dipole alignment. *Protein Science* 2, 1697-1714 (1993)

93. A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky and H. A. Scheraga: Prediction of protein conformation on the basis of a search for compact structures; test on avian pancreatic polypeptide. *Protein Science* 2, 1715-1731 (1993)

94. A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky and H. A. Scheraga: A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J Comput Chem* 18, 849-873 (1997)

95. A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, S. Oldziej and H. A. Scheraga: A united-residue force field for off-lattice protein-structure simulations. II. Parameterization of short-range interactions and determination of weights of energy terms by Z-score optimization. *J Comput Chem* 18, 874-887 (1997)

96. A. Liwo, R. Kazmierkiewicz, C. Czaplewski, M. Groth, S. Oldziej, R. J. Wawak, S. Rackovsky, M. R. Pincus, and H. A. Scheraga: United-residue force field for off-lattice protein-structure simulations; III. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials. *J Comput Chem* 19, 259-276 (1998)

97. A. Liwo, C. Czaplewski, J. Pillardy and H. A. Scheraga: Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *J Chem Phys* 115, 2323-2347 (2001)

98. J. Lee, D. R. Ripoll, C. Czaplewski, J. Pillardy, W. J. Wedemeyer and H. A. Scheraga: Optimization of parameters in macromolecular potential energy functions by conformational space annealing. *J Phys Chem B* 105, 7291-7298 (2001)

99. J. Pillardy, C. Czaplewski, A. Liwo, J. Lee, D. R. Ripoll, R. Kazmierkiewicz, S. Oldziej, W. J. Wedemeyer, K. D. Gibson, Y. A. Arnautova, J. Saunders, Y.-J. Ye and H. A. Scheraga: Recent improvements in prediction of protein structure by global optimization of a potential energy function. *Proc Natl Acad Sci USA* 98, 2329-2333 (2001)

100. A. Liwo, P. Arlukowicz, C. Czaplewski, S. Oldziej, J. Pillardy and H. A. Scheraga: A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: Application to the UNRES force field. *Proc Natl Acad Sci USA* 99, 1937-1942 (2002)

101. D. T. Jones: Progress in protein structure prediction. *Current Opinion in Structural Biology* 7, 377-387 (1997)

102. L. A. Mirny and E. I. Shakhnovich: Protein structure prediction by threading. Why it works and why it does not. *J Mol Biol* 283, 507-526 (1998)

103. A. Fersht: Structure and Mechanism in Protein Science. 536, W. H. Freeman and Co. (1999)

104. J. Lee, H. A. Scheraga and S. Rackovsky: New optimization method for conformational energy calculations on polypeptides: Conformational space annealing. *J Comput Chem* 18, 1222-1232 (1997)

105. J. Lee and H. A. Scheraga: Conformational space annealing by parallel computations: Extensive conformational search of Met-enkephalin and of the 20-residue membrane-bound portion of melittin. *Intl J of Quantum Chem* 75, 255-265 (1999)

106. R. Kazmierkiewicz, A. Liwo and H. A. Scheraga: Energy-based reconstruction of a protein backbone from its a-carbon trace by a Monte-Carlo method. *J Comput Chem* 23, 715-723 (2002)

107. R. Kazmierkiewicz, A. Liwo and H. A. Scheraga: Addition of side chains to a known backbone with defined side-chain centroids. *Biophys Chem* 100, 261-280. *Erratum: Biophys Chem* 106, 91 (2003)

108. R. Kubo: Generalized cumulant expansion method. *J Phys Soc Japan* 17, 1100-1120 (1962)

109. C. Czaplewski, A. Liwo, S. Oldziej and H. A. Scheraga: Improved conformational space annealing method to treat ß-structure with the UNRES force field and to enhance scalability of parallel implementation. *Polymer* 45, 677-686 (2004)

110. C. Czaplewski, S. Oldziej, A. Liwo and H. A. Scheraga: Prediction of the structures of proteins with the UNRES force field, including dynamic formation and breaking of disulfide bonds, *Protein Engineering, Design & Selection* 17, 29-36 (2004)

111. J. Lee, A. Liwo and H. A. Scheraga - Energy-based *de novo* protein folding by conformational space annealing and an off-lattice united-residue force field: Application to the 10-55 fragment of staphylococcal protein A and to apo calbindin D9K. *Proc Natl Acad Sci USA* 96, 2025-2030 (1999)

112. B. A. Reva, A. V. Finkelstein and J. Skolnick: What is probability of a chance prediction of a protein structure with an RMSD of 6Å? *Folding & Design* 3, 141-147 (1998)

113. A. Liwo, J. Lee, D. R. Ripoll, J. Pillardy and H. A. Scheraga: Protein structure prediction by global optimization of a potential energy function. *Proc Natl Acad Sci USA* 96, 5482-5485 (1999)

114. S. Rhee, R. G. Martin, J. L. Rosner, D. R. Davies: A novel DNA-binding motif in mar A: The first structure of an AraC family transcriptional activator. *Proc Natl Acad Sci USA* 95, 10413-10418 (1998)

115. C.A. Orengo, J. E. Bray, T. Hubbard, L. LoConte and I. Sillitoe: Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction, *Proteins: Structure. Function and Genetics* 37, S3, 149-170 (1999)

116. T. Kortemme, M. Ramirez-Alvarado and L. Serrano: Design of a 20-amino acid, three-stranded ß-sheet protein. *Science* 281, 253-256 (1998)

117. A. Kolinski, A. Godzik and J. Skolnick: A general method for the prediction of the three-dimensional structure and folding pathway of globular proteins: Application to designed helical proteins. *J Chem Phys* 98, 7420-7433 (1993)

118. C. González, G. M. Langdon, M. Bruix, A. Gálvez, E. Valdivia, M. Maquaeda and M. Rico: Bacteriocin AS -48, a microbial cyclic polypeptide structurally and functionally related to mammalian NK-lysin. *Proc Natl Acad Sci USA* 97, 11221-11226 (2000)

119. A. Liwo, P. Arlukowicz, S. Oldziej, C. Czaplewski, M. Makowski and H. A. Scheraga: Optimization of the UNRES force field by hierarchical design of the potential-energy landscape. I: Tests of the approach using simple lattice protein models. *J Phys Chem B* 108, 16918-16933 (2004)

120. S. Oldziej, A. Liwo, C. Czaplewski, J. Pillardy and H. A. Scheraga: Optimization of the UNRES force field by hierarchical design of the potential-energy landscape. II: Off-lattice tests of the method with single proteins. *J Phys Chem B* 108, 16934-16949 (2004)

121. S. Oldziej, J. Lagiewka, A. Liwo, C. Czaplewski, M. Chinchio, M. Nanias and H. A. Scheraga: Optimization of the UNRES force field by hierarchical design of the potential-energy landscape: III. Use of many proteins in optimization. *J Phys Chem B* 108, 16950-16959 (2004)

122. J. A. Saunders and H. A. Scheraga: Ab initio structure prediction of two a-helical oligomers with a multiple-chain united residue force field and global search. *Biopolymers* 68, 300-317 (2003)

123. J. A. Saunders and H. A. Scheraga: Challenges in structure prediction of oligomeric proteins at the united-residue level: searching the multiple-chain energy landscape with CSA and CFMC. *Biopolymers* 68, 318-332 (2003)

124. Y.-J. Ye and H. A. Scheraga: Kinetics of protein folding, in "Slow Dynamics in Complex Systems: Eighth Tohwa University International Symposium", Eds. M. Tokuyama and I. Oppenheim. AIP Conference Proceedings. *Amer Inst Phys* 469, 452-475 (1999)

125. Y.-J. Ye, D.R. Ripoll and H. A. Scheraga: Kinetics of cooperative protein folding involving two separate conformational families. *Comput and Theor Polymer Sci* 9, 359-370 (1999)

126. R. Elber, A. Ghosh and A. Cárdenas: Long-time dynamics of complex systems. *Acc Chem Res* 35, 396-403 (2002)

127. S. P. Bottomley, A.G. Popplewell, M. Scawen, T. Wan, B. J. Sutton and M. G. Gore: The stability and unfolding of an IgG-binding protein based upon the B-domain of protein A from staphylococcus aureus probed by tryptophan substitution and fluorescence spectroscopy. *Protein Eng* 7, 1463-1470 (1994)

128. Y. Bai, A. Karimi, H. J. Dyson and P. E. Wright: Absence of a stable intermediate on the folding pathway of protein A. *Protein Sci* 6, 1449-1457 (1997)

Key Words: protein folding, empirical force fields, global optimization, folding pathways

Send correspondence to: Dr H. A. Scheraga, Baker Laboratory of Chemistry, Cornell University, Ithaca, NY 14853, Tel: 607-255-4034, Fax: 607-254-4700, E-mail: has5@cornell.edu