
Journal of Global Optimization **25:** 113–140, 2003.

Ab initio Tertiary Structure Prediction of Proteins

J. L. KLEPEIS and C.A. FLOUDAS*
Department of Chemical Engineering, Princeton University, Princeton, NJ 08544-5263, U.S.A
*(*Author for correspondence. Tel.: +1-609-258-4595; Fax: +1-609-258-0211; e-mail: floudas@titan.princeton.edu)*

April 12, 2002

Abstract. A daunting challenge in the area of computational biology has been to develop a method to theoretically predict the correct three-dimensional structure of a protein given its linear amino acid sequence. The ability to surmount this challenge, which is known as the protein folding problem, has tremendous implications. We introduce a novel ab initio approach for the protein folding problem. The accurate prediction of the three-dimensional structure of a protein relies on both the mathematical model used to mimic the protein system and the technique used to identify the correct structure. The models employed are based solely on first principles, as opposed to the myriad of techniques relying on information from statistical databases. The framework integrates our recently proposed methods for the prediction of secondary structural features including helices and strands, as well as β-sheet and disulfide bridge formation. The final stage of the approach, which culminates in the tertiary structure prediction of a protein, utilizes search techniques grounded on the foundations of deterministic global optimization, powerful methods which can potentially guarantee the correct identification of a protein's structure. The performance of the approach is illustrated with bovine pancreatic trypsin inhibitor protein and the immunoglobulin binding domain of protein G.

Key words: Protein folding, Tertiary structure prediction, Secondary structure, Global optimization

1. Introduction

Proteins serve as vital components in our cellular makeup and perform many biological functions that are essential for sustaining life. An important feature which determines the functionality of a protein is the form of its three-dimensional structure. The structure, in turn, is related to the protein sequences encoded by our genes, and these sequences have recently been identified as part of the data from the human genome project. Therefore, a logical undertaking upon completion of the human genome project, and an important step in understanding and treating disease, is to develop a method to predict the structure of a protein given its sequence information.

The difficulty in addressing the protein folding problem arises from the complexity inherent to the intricate balance of molecular forces which define the native three-dimensional structure of the system. Experimental observations have revealed the fact that many proteins fold spontaneously from random disordered

states into compact states of unique shape. However, the mechanisms that govern this transformation are not yet fully understood.

Accurate prediction of the three-dimensional structure of a protein relies on both the mathematical model used to mimic the protein system and the technique used to identify the correct structure. This work represents the final stage in a novel approach for *ab initio* prediction of the three dimensional structures of proteins. In contrast to database driven predictions, the initial stages of this approach provide information on both secondary and tertiary structural features using only the amino acid sequence. The methods rely on a coupled decomposition, combinatorial optimization and global optimization scheme to identify dominant structural elements. In the final stage, this information is then combined to formulate a global optimization problem for the full system, whose solution provides the overall tertiary fold of the protein.

The first stage of the approach focuses on secondary structure prediction of α-helical segments. This is accomplished through detailed atomistic level modeling of overlapping subsequences of the overall protein sequence and free energy calculations. For each subsequence, global optimization is used to identify an ensemble of low energy structures along with the global minimum energy conformation. Analysis of these results provides a means to identify the potential for α-helix formation (Klepeis and Floudas, 2002a).

The positions of additional secondary structural elements, including β-strand conformations, are determined through the analysis of residue properties, including hydrophobic propensities. Residue classifications are then used to formulate a problem to predict the formation of ordered structural features, such as parallel and antiparallel β-sheets. This formulation results in a set of integer linear programming (ILP) problems, which can be solved to global optimality to identify a set of optimal hydrophobic contacts. Solutions to these ILP problems represent potential β-sheet configurations for the overall protein. The formation of disulfide bonding pairs can also be identified within the context of the ILP model (Klepeis and Floudas, 2002b).

The final stage of the approach couples the preceding information to predict the overall tertiary fold of the protein. The positions of both α and β secondary elements and disulfide bonding pairs are first used to derive distance and angle restraints. This leads to a sparse system of restraints which can be treated as a constrained global optimization problem. The formulation incorporates detailed atomistic level energy modeling so that the energetically most stable conformation satisfying the imposed constraints is identified.

These modeling and optimization components have been combined into a single approach, ASTRO-FOLD, which predicts the three dimensional structures of proteins given only their amino acid sequence. The overall approach, a schematic of which is given in Figure 1, has been applied to several protein sequences, including the 58 residue bovine pancreatic trypsin inhibitor protein and the 56 residue im-

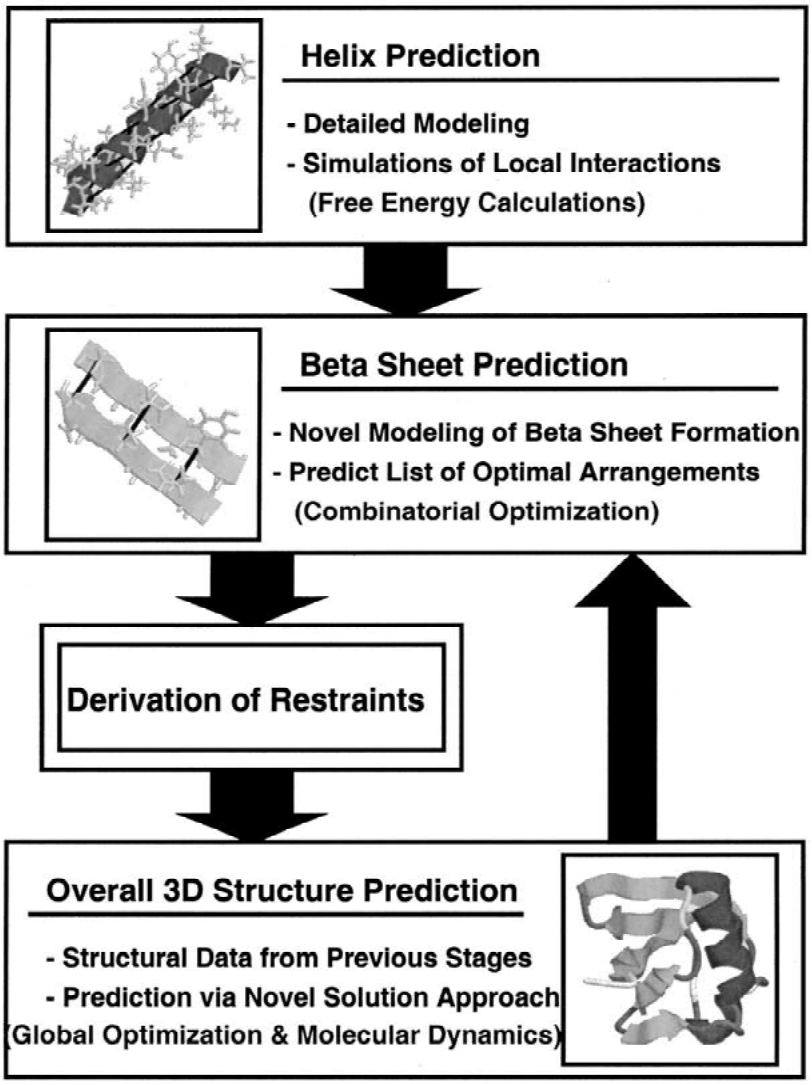

Figure 1. Overall schematic for prediction of native three-dimensional structures of proteins using ASTRO-FOLD.

munoglobulin binding protein of protein G. These results indicate the possibility for significant advances in the theoretical treatment of the protein folding problem.

2. Physical understanding of protein Ffolding

An important question regarding the prediction of the native folded state of a protein is how the formation of secondary and tertiary structure proceeds. Two common viewpoints provide competing explanations to this question. The classical

opinion regards folding as hierarchic, implying that the process is initiated by fast formation of secondary structural elements, followed by the slower arrangement of the tertiary fold. The opposing perspective is based on the idea of a hydrophobic collapse, which suggests that tertiary and secondary features form concurrently.

Inherent to the hierarchical view of protein folding is the dominant role of local forces in determining the formation of secondary structure. These local forces denote those interactions between neighboring residues, rather than nonlocal forces that may arise during tertiary structure formation. In other words, local sequence information should be sufficient to predict native secondary structure if folding is hierarchic. In considering the local prediction of secondary structure elements, such as α-helices, β-strands and turns, most methods rely on statistical treatments (Munoz and Serrano, 1994). More recent work has led to the proposal of a physical theory for secondary structure formation based on local interactions and sterics (Baldwin and Rose, 1999a,1999b; Srinivasan and Rose, 1999) The basis for this theory hinges on the role of intrinsic propensities for backbone conformations and backbone hydrogen bonding.

The alternative perspective stresses the importance of the hydrophobic collapse rather than local propensities in determining a protein's fold. In this view, hydrophobic forces drive the collapse through the desolvation of side chains. It is believed that these non-local side chain interactions influence the formation of tertiary as well as secondary structural elements (Dill, 1999). In addition, these ideas suggest that simple side chain models of protein folding may be sufficient to predict folding behavior.

For both cases experimental evidence has been produced to support the underlying claims. For example, kinetic studies have shown that elements of secondary structure common to the native fold are able to form before substantial tertiary structure arrangement. The boundaries of helical structure can also be identified through local sequence information, implying that local interactions dominate helix formation. In addition, fragments of longer protein sequences can form native-like folds in absence of long range interactions (Baldwin, 1995). On the other hand, support for non-hierarchical folding through a hydrophobic collapse includes experiments showing that protein folds are less affected by mutations on their surfaces than in their hydrophobic cores (Lim and Sauer, 1991). In addition, hydrophobic collapse, like secondary structure formation, occurs rapidly in certain cases (Chan et al., 1997). Other results, such as the formation of β-sheet folds through α-helical intermediates (Hamada et al., 1996), imply that secondary units are not preassembled and can be driven by tertiary structure formation.

3. Secondary structure prediction

Secondary structure prediction is an important precursor in tackling the overall protein folding problem, and many methods have been developed in an attempt to accurately predict the location of α-helices and β-strands. The most successful

methods rely on homology modeling or multiple sequence alignments to predict secondary structure using only the amino acid sequence. If the databases of experimental structures contain significantly similar (homologous) sequences to the target sequence, then local conformation patterns, such as α-helices and β-strands, can be predicted with accuracy that in certain cases can exceed 70 percent. However, many protein sequences do not possess known structural homologues, which causes a significant decrease in prediction accuracy.

3.1. HELIX PREDICTION

It is interesting to note that simulations of a hydrophobic collapse through side chain models fail to predict the formation of α-helices. This indicates that simplified models for protein folding may not be sufficient because they lack a full structural and energetic description of secondary structure formation. Other methods, such as those based on a statistical treatment for helix determination, as referred to above, have been effective, but lack a true physical basis.

Recently, a method has been presented in which the physical principles of hierarchical folding are used for the prediction of α-helices in protein systems (Klepeis and Floudas, 2002a). The support for this approach for α-helix determination is based on observations that native like segments of helical secondary structure form rapidly. The ability for helices to overcome Levinthal's paradox suggests that α-helix formation can occur during the earliest stages of protein folding. Such a mechanism for the helix-coil transition is based on local interactions which induce nucleation and propagation of the helix.

An important component of this approach is that some information regarding helix formation is retained locally, which is evidenced by experimental observations regarding the strong nucleation characteristics of helices. To capture local interactions and the unique positioning of each residue in the overall protein, the protein sequence is decomposed into overlapping oligopeptides. The analysis also involves detailed atomistic level modeling, and the refinement of helix propensities according to polarization and ionization energies calculated through the solution of the Poisson Boltzmann equation, which eliminates approximations based on force field electrostatics. The end result is the prediction of helical segments according to the average helix propensity assigned to each residue.

The overall approach for the ab initio prediction of helical segments in polypeptides is based on the key ideas of (i) partitioning the sequence of aminoacids into oligopeptides (e.g., pentapeptides, heptapeptides) such that consecutive oligopeptides are shifted by one aminoacid. (ii) atomistic level modeling of all appropriate interactions for each oligopeptide using the ECEPP/3 force field; (iii) generation of an ensemble of low energy conformations for each oligopeptide using global optimization based approaches (e.g., αBB, CSA); (iv) incorporation of the entropic contributions and free energy calculations for each oligopeptide; (v) calculations of the contributions to free energy due to the formation of cavity for

selected oligopeptides; (vi) calculations of the solvation contribution to free energy using the nonlinear Poisson–Boltzmann equation for selected oligopeptides; (vii) calculations of the ionization contribution to free energy using the nonlinear Poisson–Boltzmann equation for selected oligopeptides; (viii) calculation of equilibrium occupational probabilities for the helical clusters based on the free energies of the oligopeptides; and (ix) classification of residues as helical according to average propensities for each residue as calculated by the equilibrium occupational probabilities for the helical clusters.

The approach has been applied the location of α and $3-_{10}$ helices for a variety of proteins which have been studied both experimentally and through simulation (Klepeis and Floudas, 2002a). The results provided by this approach for bovine pancreatic trypsin inhibitor protein and the immunoglobulin binding domain of Protein G are used as input for the overall tertiary structure prediction of the system. For the case of bovine pancreatic trypsin inhibitor, helical regions are predicted between residues 2–5 and 47–54, which agrees well with the experimental observations of helices between 3–6 and 46–55. For the immunoglobulin binding domain of Protein G, one α-helix domain is assigned between residues 23–34, almost exactly matching position of the helix in the experimental structure.

3.2. β-STRUCTURE PREDICTION

A major hindrance to the accurate prediction of the tertiary structure of proteins has been the correct identification of β-structure. A number of methods provide relatively effective means for anticipating the positions of helical segments; however, many of these methods have been based on the use of statistical databases and pattern recognition algorithms for three-state (helix, extended, coil) secondary structure prediction. In stark contrast to the helix predictions, the statistical methods regularly fail in the prediction of β-strands.

To circumvent these limitations, a novel methodology for the prediction of β-strand and β-sheet conformations, as well as disulfide bridge arrangements has been developed (Klepeis and Floudas, 2002b). In this approach, a simplified model is developed according to residue properties, including hydrophobic propensities, which can be derived from experimental or purely computational information. Residue classifications are used to formulate a problem to predict the formation of ordered structural features, such as parallel and antiparallel β-sheets. This formulation results in a set of integer linear programming (ILP) problems, which can be solved to global optimality to identify the optimal set of hydrophobic contacts. Solutions to these (ILP) problems represent potential β-sheet configurations for the overall protein. The formation of disulfide bonding pairs can be identified within the context of the (ILP) model.

The proposed approach for the prediction of antiparallel β-sheets, parallel β-sheets and disulfide bridges borrows key concepts from a mathematical framework developed in the area of process synthesis of chemical systems (Floudas, 1995).

The approach is based on the idea that β-structure formation relies on a hydrophobic driving force. To model this hydrophobic force, it is necessary to predict contacts between hydrophobic residues.

The first important component is the postulation of a β-strand superstructure that encompasses all alternative β-strand arrangements of interest. It is important to emphasize that the superstructure may include more β-strands than needed. That is, it may postulate the existence of a β-strand which may eventually not be selected to participate in the β-sheet topology, and therefore not exist as a β-strand.

The second key component involves the development of a single mathematical model to describe the topology of the postulated superstructure. This model includes binary variables representing the existence or not of the β-strands and binary variables denoting the connectivity of the postulated β-strands. In addition, several constraint sets are delineated in the model so as to represent the antiparallel and parallel arrangements, the physically consistent structures and the disulfide bridges. The main concept in the model derivation relies upon the potential contacts between pairs of hydrophobic amino acids, and the objective function aims at maximizing the hydrophobic–hydrophobic contact energy. The proposed model is an Integer-Linear Programming (ILP) model.

The third component of the proposed framework is the solution of the resulting mathematical model that extracts from the postulated β-strand superstructure the globally optimal solutions of (a) the contacts of hydrophobic residues, (b) the existence of β-strands and their arrangements to form β-sheets, and (c) the disulfide bridge configuration. It is important to emphasize that given the nature of the (ILP) model, a rank ordered list of the second best, third best, etc. solutions can be generated along with the globally optimal solution.

The most important aspect of the approach is not the accurate prediction of potential β-strands, but that β-sheet, including disulfide bridge, topologies are identified. The approach has been used to predict β-strand topologies for a number of protein systems, and the results for bovine pancreatic inhibitor protein and the immunoglobulin binding domain of protein G are in included in this work for the tertiary structure prediction of these proteins. Specifically, for bovine pancreatic trypsin inhibitor the approach provides a match for β-strands between residues 17–23 and 29–35, as well as an additional contact between residues 44–46 and 20–22. For the immunoglobulin binding domain of protein G, β-sheets form between strands defined by residues 43–45 and 51–55, as well strands defined by residues 1–7 and 16–21. An additional match is predicted to occur between the first and last β-strands, in accordance with parallel β-sheet formation.

4. Derivation of restraints

4.1. HELIX AND β-RESTRAINTS

For those residues predicted to assume α-helical and β-strand conformations, dihedral angle bounds are assigned according to the values given in Table 1. In addition, for α-helices, C^{α}–C^{α} distances can be restrained between each pair of i and $i+3$ residues, in anticipation of the formation of the α-helix hydrogen bond network. In a similar fashion, C^{α}–C^{α} restraints can be developed for residues in opposing strands of a β-sheet fold, so that hydrogen bond formation between strands is enforced. Unlike the helix-based restraints, these restraints are not based on local interactions; instead, the restraints reflect tertiary structure formation between opposing strands in the β-sheet network. The β-strand restraints include both hydrophobic residues and intervening residues; more specifically, between the turn and the full extent of the β-sheet. The corresponding upper and lower distance bounds are given in Table 2. Finally, distance restraints are included for those cystine residues participating in a disulfide bridge network. In this case, sulfur atoms of the opposing cystine residue are constrained between 2.01 and 2.03 *Å*.

Table 1. Dihedral angle bounds, lower and upper, for α-helix and β-strand residues

Conformer	ϕ^L	ϕ^U	ψ^L	ψ^U
α	-85	-55	-50	-10
β	-155	-75	110	180

Table 2. C^{α}-C^{α} distance bounds, lower and upper, for α-helix and β-strand residues

Conformer	d^L	d^U
α	5.50	6.50
β	4.50	6.50

4.2. LOOP RESTRAINTS

Additional restraints can be generated through analysis of the unassigned residues, that is, those not assigned to helical or strand regions in the protein sequence. Specifically, fragments of unassigned residues between two consecutive secondary structures comprise a set of candidate loop segments. Two methods have been

$$
\begin{aligned}
E_{\text{ECEPP/3}} = & \sum_{(ij)\in \text{ES}} \frac{q_i q_j}{r_{ij}} && \text{(Electrostatic)} \\
+ & \sum_{(ij)\in \text{NB}} F_{ij} \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}} && \text{(Nonbonded)} \\
+ & \sum_{(ij)\in \text{HX}} \frac{A'_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{10}} && \text{(Hydrogen bonded)} \\
+ & \sum_{k\in \text{TOR}} \left(\frac{E_{o,k}}{2}\right)(1+c_k \cos n_k \theta_k) && \text{(Torsional)}
\end{aligned}
$$

Figure 2. Potential energy terms in ECEPP/3 force field. r_{ij} refers to the interatomic distance of the atomic pair (ij). Q_i and Q_j are dipole parameters for the respective atoms, in which the dielectric constant of 2 has been incorporated. F_{ij} is set equal to 0.5 for 1–4 interactions and 1.0 for 1–5 and higher interactions. A_{ij}, C_{ij}, A'_{ij} and B_{ij} are nonbonded and hydrogen bonded parameters specific to the atomic pair. $E_{o,k}$ are parameters corresponding to torsional barrier energies for a given dihedral angle. θ_k represents any dihedral angle. c_k takes the values -1,1, and n_k refers to the symmetry type for the particular dihedral angle.

developed for the derivation of these restraints. The first method proceeds in a manner similar to that of the prediction of helical segments in that a series of free energy calculation for overlapping oligopeptides is performed. The second method involves the simulation of the entire loop fragment. The form of the derived restraints include tightened dihedral angle bounds (ϕ^L and ϕ^U) for those residues connecting consecutive elements of secondary structure.

For both approaches the derivation of restraints involves the sampling of the energy surface through global minimization and the generation of a low energy ensemble. The basic formulation involves:

$$
\begin{aligned}
\min_{\phi} \quad & E_{\text{forcefield}}(\phi) \qquad (1) \\
\text{subject to} \quad & \phi_i^L \leqslant \phi_i \leqslant \phi_i^U, i = 1, \ldots, N_\phi.
\end{aligned}
$$

Here the ϕ represent the variables used to describe a protein conformation in the torsion angle space, while $\phi^{\mathbf{L}}$ and $\phi^{\mathbf{U}}$ indicate the lower and upper bounds on these variables (which include both backbone and side chain degrees of freedom). The energy function, $E_{\text{forcefield}}(\phi)$ is based on the atomistic level ECEPP/3 force field, which is described in Figure 2 (Némethy et al., 1992). The basic components include electrostatic terms, 6–12 based van der Waals potential for non-bonded interactions, modified 10–12 van der Waals potential for possible hydrogen-bonded interactions, and torsional terms. The detailed energy modeling greatly increases the complexity of the objective function. It should also be noted that the transformation from Cartesian to internal coordinate space results in highly nonlinear functions. That is there is not a one-to-one correspondence between distances and internal coordinates. The advantage for working in dihedral angle space is that the variable set decreases, with the disadvantage being the increased nonconvexity of the energy hypersurface.

In order to solve the formulation given by Equation (1) to provide ensembles of low energy conformers, powerful search techniques classified within the realm of global optimization must be utilized. Although many such methods have been developed, the major limitation is that the majority of the methods depend strongly on heuristics and initial point selection. To circumvent these difficulties, the application of deterministically based global optimization approaches is required. One such method, the αBB global optimization approach (Androulakis et al., 1995; Adjiman et al., 1996,1997,1998a,b) has been extended to identifying global minimum energy conformations of peptides. The development of this branch and bound method was motivated by the need for an algorithm that could guarantee convergence to the global minimum of nonlinear optimization problems with twice-differentiable functions (Floudas, 1997,2000). The application of the αBB to the minimization of potential energy functions was first introduced for microclusters (Maranas and Floudas, 1992,1993), and small acyclic molecules (Maranas and Floudas, 1994a,b). The αBB approach has also been applied to general constrained optimization problems (Androulakis et al., 1995; Adjiman et al., 1996,1998a,b) In more recent work, the algorithm has been shown to be successful for isolated peptide systems using the ECEPP/3 potential energy model (Maranas et al., 1996; Androulakis et al., 1997), and including several solvation models (Klepeis et al., 1998; Klepeis and Floudas, 1999). αBB based global optimization techniques have also been applied to NMR type structure prediction problems (Eyrich et al., 1999; Klepeis et al., 1999; Standley et al., 1999).

The αBB global optimization approach effectively brackets the global minimum by developing converging sequences of lower and upper bounds. These bounds are refined by iteratively partitioning the initial domain. Upper bounds on the global minimum are obtained by local minimizations of the original nonconvex problem. Lower bounds belong to the set of solutions of the convex lower bounding problems, which are constructed by augmenting the objective and constraint functions through the addition of separable quadratic terms. The lower bounding formulation can be expressed in the following manner:

$$\begin{aligned} \min_{\phi} \quad & L_{\text{forcefield}}(\phi), \\ \text{subject to} \quad & \phi_i^L \leqslant \phi_i \leqslant \phi_i^U, i = 1, \ldots, N_\phi. \end{aligned} \tag{2}$$

In this formulation, variable bounds are specific to the subdomain for which the lower bounding functions are constructed. $L_{\text{forcefield}}$ refers to the convex representation of the objective function, as given by:

$$L_{\text{forcefield}} = E_{\text{forcefield}} + \sum_{i=1}^{N_\phi} \alpha_{\phi_i} \left(\phi_i^L - \phi_i\right)\left(\phi_i^U - \phi_i\right). \tag{3}$$

The α parameters represent nonnegative parameters which must be greater or equal to the negative one–half of the minimum eigenvalue of the Hessian of $E_{\text{forcefield}}$ over

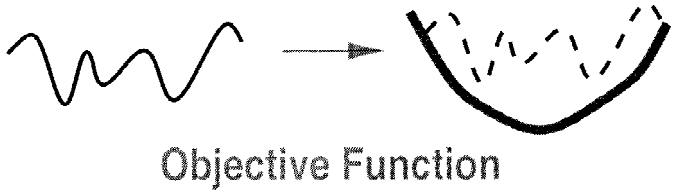

Figure 3. Convex underestimator constructed for one-dimensional nonconvex objective function.

the defined domain. Rigorous bounds on the α parameters can be obtained through a variety of approaches (Maranas and Floudas, 1994a; Adjiman and Floudas, 1996; Adjiman et al., 1998a,b; Hertz et al., 1999). The overall effect of these terms is to overpower the nonconvexities of the original terms by adding the value of 2α to the eigenvalues of the Hessian of $E_{\text{forcefield}}$. An illustration of the convexification of a nonconvex objective function is given in Figure 3. The αBB approach can also be applied to general formulations involving nonlinear constraint sets. For example, the use of nonlinear distance constraints requires a reformulation of the problem given in Equation (1), the solution of which will be detailed in section Section 5.

Once solutions for the upper and lower bounding problems have been established, the next step is to modify these problems for the next iteration. This is accomplished by successively partitioning the initial domain into smaller subdomains. For the protein conformation problems, it has been found that an effective partitioning strategy involves bisecting the same variable dimension across all nodes at a given level. In order to ensure non-decreasing lower bounds, the hyper-rectangle to be bisected is chosen by selecting the region which contains the infimum of the minima of lower bounds. A non–increasing sequence for the upper bound is found by solving the nonconvex problem locally and selecting it to be the minimum among all conformers in the upper bound list. If the single minimum of $L_{\text{forcefield}}$ for any hyper-rectangle is greater than the current upper bound, the global minimum cannot exist within this region and the entire subdomain can be deleted from the list of searchable regions (fathoming step). The computational requirement of the αBB algorithm depends on the number of variables (global) on which branching occurs.

The use of the αBB method is also amenable to the integration of other stochastic or heuristic search techniques for enhancing and improving the identification of low energy conformations. In other words, the solution of the upper bounding problem (i.e., the original nonconvex problem) is not limited to the use of nonlinear local minimization techniques. For the problems related to the derivation of loop restraints, a particularly successful marriage is the use of the Conformational Space Annealing (CSA) algorithm (Lee et al., 1997) as an upper bounding solver within the αBB framework (Klepeis et al., 2002).

The end result of this procedure is a set of restraints for those residues connecting consecutive elements of secondary structure. In particular, these restraints

take the form of reduced ϕ and ψ domains for the loop residues, which are extracted from the set of low free energy conformers identified for the oligopeptides representing these segments. For smaller loop segments, and for those segments connecting β-sheet topology, the reduced dihedral angle bounds are relatively tight, thus focusing the search for the overall three dimensional structure.

5. Tertiary Structure Prediction

The final stage of the approach involves the prediction of the tertiary structure of a full protein sequence. The problem formulation is based on the development of atomic distance and dihedral angle restraints derived from the α-helix, β-sheet and loop prediction results, as detailed in Section 4. In its final form, the problem requires the use of constrained nonlinear global optimization techniques. In this work, a combination of the deterministically based αBB global optimization approach and molecular dynamics in torsion angle space (TAD) is used to solve this problem (Klepeis et al., 1999).

The restrained global minimization problem is formulated as an unconstrained minimization with a hybrid energy function :

$$E = E_{\text{forcefield}} + W_{\text{res}} E_{\text{res}}. \tag{4}$$

The energy, E, specified by this target function includes a chemical description of the protein conformation through the use of a force field, $E_{\text{forcefield}}$. These force field potentials are typically much simpler representations of all atom force fields. The distance and dihedral angle restraints appear as penalty terms, E_{res}, with weights, W_{res}, that should be driven to zero.

When expanded, the second term of Equation (4) becomes:

$$E_{\text{res}} = E_{\text{distance}} + E_{\text{dihedral}}. \tag{5}$$

Here E_{distance} and E_{dihedral} represent the violation energies based on the distance and dihedral angle restraints, respectively. These functions can take several forms, although a simple square well potential is commonly used, and includes a summation over both upper and lower distance violations. For example, $E_{\text{distance}} = E^{\text{upp}}_{\text{distance}} + E^{\text{low}}_{\text{distance}}$. For upper distance restraints:

$$E^{\text{upp}}_{\text{distance}} = \sum_j \begin{cases} A_j (d_j - d_j^{\text{upp}})^2 & \text{if } d_j > d_j^{\text{upp}}, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The squared violation energy is considered only when the calculated distance d_j exceeds the upper reference distance d_j^{upp}. This squared violation is multiplied by a weighting factor A_j. Similarly, violations are calculated for those distances that deviate beyond a lower distance limit d_j^{low} :

$$E^{\text{low}}_{\text{distance}} = \sum_j \begin{cases} A_j (d_j - d_j^{\text{low}})^2 & \text{if } d_j < d_j^{\text{low}}, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

For dihedral angle restraints the functional form is similar to that of Equations (6) and (7) in that the total violation, E_{dihedral}, is a sum over upper and lower violations (i.e., $E_{\text{dihedral}} = E^{\text{upp}}_{\text{dihedral}} + E^{\text{low}}_{\text{dihedral}}$). A dihedral angle, ω_j, can be restrained through a quadratic square well potential using upper (ω_j^{upp}) and lower (ω_j^{low}) bounds on the variable values. However, due to the periodic nature of these variables, a scaling parameter must be incorporated to capture the symmetry of the system. The full periodic region is centered on the region defined by the allowable bounds so that all transformed values will lie in the domain defined by $[\omega_j^{\text{low}} - \Delta H_{\omega_j}, \omega_j^{\text{upp}} + \Delta H_{\omega_j}]$, where ΔH_{ω_j} is equal to half the excluded range of dihedral angle values (i.e., $\Delta H_{\omega_j} = \pi - (\omega_j^{\text{upp}} - \omega_j^{\text{low}})/2$). This results in the following set of equations:

$$E^{\text{upp}}_{\text{dihedral}} = \sum_j \begin{cases} A_j \left(1 - 2\left[\frac{\omega_j - \omega_j^{\text{upp}}}{2\pi - (\omega_j^{\text{upp}} - \omega_j^{\text{low}})}\right]^2\right)(\omega_j - \omega_j^{\text{upp}})^2 & \text{if } \omega_j > \omega_j^{\text{upp}}, \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

$$E^{\text{low}}_{\text{dihedral}} = \sum_j \begin{cases} A_j \left(1 - 2\left[\frac{\omega_j - \omega_j^{\text{low}}}{2\pi - (\omega_j^{\text{upp}} - \omega_j^{\text{low}})}\right]^2\right)(\omega_j - \omega_j^{\text{low}})^2 & \text{if } \omega_j < \omega_j^{\text{low}}, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Since the sets of restraints provided by the initial stages of the approach are relatively sparse, a detailed atomistic force field is employed in lieu of the more typically simplistic potentials. Specifically, the ECEPP/3 force field, as depicted in Figure 2, is again utilized. With these modifications, the objective function of Equation (4) becomes :

$$E_{\text{D}} = E_{\text{distance}} + E_{\text{dihedral}} + E_{\text{ECEPP/3}}. \tag{10}$$

The usual method for solving the unconstrained global optimization problem presented by Equation (10), which is similar in form to those formulations for NMR structure prediction, is to explore the conformational space using a combination of simulated annealing and molecular dynamics. These techniques are stochastic in nature and generally require the selection of a set of initial points to be optimized. The conformers are then grouped according to their evaluation, from which a sample of approximately 100 structures are selected for further analysis. From this group a smaller subset of 20–50 conformers are used to characterize the system. In terms of global optimization, simulated annealing is used because it provides a means for escaping local minimum energy wells and broadening the search.

The advances in this work include the transformation of the mathematical formulation of the general unconstrained problem to a constrained global optimization

problem. This reformulation removes both E_{dihedral} and E_{distance} from the target function, leaving only $E_{\text{forcefield}}$:

$$\min_{\phi} \quad E_{\text{ECEPP/3}}, \tag{11}$$

$$\text{subject to} \quad E_l^{\text{distance}}(\phi) \leqslant E_l^{\text{ref}} \quad l = 1, \ldots, N_{\text{CON}},$$
$$\phi_i^L \leqslant \phi_i \leqslant \phi_i^U, \; i = 1, \ldots, N_\phi.$$

As before $i = 1, \ldots, N_\phi$ corresponds to the set of dihedral angles, ϕ_i, with ϕ_i^L and ϕ_i^U representing lower and upper bounds on these dihedral angles. In general, the lower and upper bounds for these variables are set to -π and π, although appropriate bounds, as derived above, are also suitable. E_l^{ref} are reference parameters for the N_{CON} constraints. The set of constraints are completely general, and represent either the full combination of distance restraints or smaller subsets of the define distance restraints. A physically appealing procedure is to define a single restraint for each element of structure. That is, for each helix, the set of hydrogen bonding distances that define the helix can be formulated as a single restraint and controlled individually. In a similar fashion for each β-sheet match, the set of hydrogen bond distances can be combined to form a single constraint. In this way the maximum and average violation for each structural element can be controlled separately. A constraint including all distance can also be included to limit the violation of the total structure.

The constraints, through reduction of the feasible search space, serve two important purposes: (1) attempt to correct any deficiencies of the energy model, and (2) focus the efforts of the global optimization algorithm. Through the use of this constrained optimization approach, the dihedral angle bounds are implicitly included as box constraints, while distance restraints are treated explicitly.

As alluded to in Section 4, the general nonconvex constrained problem is solved via the αBB global optimization approach. As before, a converging sequence of upper and lower bounds is generated, with the upper bounds on the global minimum are obtained by local minimizations of the original nonconvex problem. Lower bounds belong to the set of solutions of the convex lower bounding problems, which are constructed by augmenting the objective and constraint functions through the addition of separable quadratic terms. The formulation for the lower bounding function becomes:

$$\min_{\phi} \quad L_{\text{forcefield}}(\phi), \tag{12}$$

$$\text{subject to} \quad L_l^{\text{distance}}(\phi) \leqslant E_l^{\text{ref}} \; l = 1, \ldots, N_{\text{CON}},$$
$$\phi_i^L \leqslant \phi_i \leqslant \phi_i^U, \; i = 1, \ldots, N_\phi.$$

The formulation is similar to that given by Equation (3) with an additional set of equations for the distance restraints. L_l^{distance} denotes the convex relaxation of these

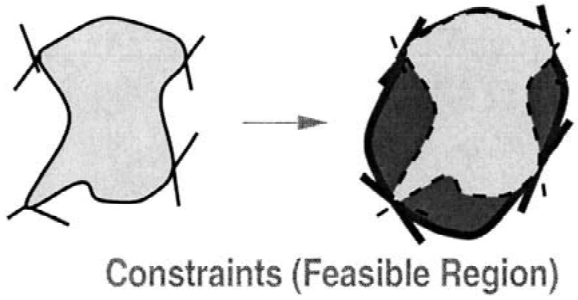

Figure 4. Convex underestimator constructed for two dimensional nonconvex constraint set.

inequality constraints as given by:

$$L_l^{\text{distance}} = E_l^{\text{distance}} + \sum_{i=1}^{N_\phi} \alpha_{\phi_i,l}^{\text{distance}} \left(\phi_i^L - \phi_i\right)\left(\phi_i^U - \phi_i\right). \tag{13}$$

An illustration of the convexification of a nonconvex constraint set is given in Figure 4, as constructed by overpowering the nonconvexities of the original function through the addition of 2α to the eigenvalues of the Hessian of $E_{\text{forcefield}}$.

The highly nonlinear form of the potential energy function coupled with the nonconvexities of the constraints substantially increases the difficulty in identifying low energy feasible points for the αBB approach. On the other hand, although TAD methods perform well for strongly determined systems, such methods are much more ambiguous for systems displaying sparse sets of restraints. To alleviate these difficulties a relatively fast TAD simulation is implemented as a preprocessing step to the local minimization of the upper bounding problem. As a result, the performance of the αBB approach is improved significantly through the rapid determination of good approximations to the global minimum energy.

5.1. TORSION ANGLE DYNAMICS

Standard unconstrained molecular dynamics simulations have been used extensively to model the folding and unfolding of protein systems (Caves et al., 1998; Daggett et al., 1998; Duan and Kollman, 1998). In addition, several methods for NMR structure calculation have been based on molecular dynamics in Cartesian space (Brünger, 1992). Torsion angle dynamics differs from traditional molecular dynamics in that bond lengths and bond angles are fixed at equilibrium values, thereby allowing for a transformation from the Cartesian to the internal coordinate system. The constraints on the systems also dampen high frequency motions, which permits the use of longer time steps during the numerical integration of the equations of motion. The use of TAD in place of conventional MD has been found to improve the efficiency of NMR structure prediction (Rice and Brünger, 1994; Güntert et al., 1997).

A major disadvantage for employing TAD in place of Cartesian MD is that the equations of motion become much more complex for the constrained system. For unconstrained Cartesian MD the accelerations of the atoms can be calculated independently due to the decoupled nature of the equations of motion. The addition of constraints to the Cartesian system transforms the equations from a system of ODEs to a system of differential algebraic equations (DAEs). The alternative to solving this system of DAEs is to transform the equations of motion to the internal coordinate reference frame. In this case, the solution of a linear matrix equation in each time step is required, which, due to the highly coupled structure of the equations, scales as a cubic function of the number of degrees of freedom (torsion angles). To avoid the potentially prohibitive computational cost required for the solution of the equations of motion, a fast recursive algorithm, which scales linearly with the number of torsion angles, was implemented. The algorithm is based on spatial operator algebra which has been used to simulate the dynamics of astronautical and robotic equipment (Jain et al., 1993), the details of which are given in the Appendix.

5.2. ALGORITHMIC STEPS

The algorithmic steps for the constrained αBB approach can be generalized to any force field model and routine for local minimization of constrained optimization problems. In this work, the αBB approach is interfaced with PACK (Scheraga, 1996) and NPSOL (Gill et al., 1986). PACK is used to transform to and from Cartesian and internal coordinate systems, which is needed to obtain function and gradient contributions for the ECEPP/3 force field and the distance constraint equations. NPSOL is a local nonlinear optimization solver that is used to locally solve the constrained upper and lower bounding problems in each subdomain.

The implementation is composed of two basic phases: *initialization* and *computation*. The basic steps of the *initialization* phase are as follows :

(1) Choose the set of global variables. Since the bounds on these variables will be refined during the course of global optimization, they should be selected according to their influence on the structure of the molecule. In this work (and in general) the ϕ and ψ (backbone) dihedral angles provide the largest structural variability, and are chosen as the global variable set. The remaining dihedral angles, which generally describe side chain configurations, are treated as local variables.

(2) Set upper and lower bounds on all dihedral angles (variables). If information is not available for a given dihedral angle, the variable bounds are set to $[-\pi, \pi]$ or to some symmetry based region. Since a constrained local optimization solver is used, these box constraints are strictly enforced.

(3) Identify the set of derived distance restraints to be used in the constraints. Although the formulation can handle multiple constraints, distance restraints were included as one constraint ($N_{\mathrm{CON}} = 1$) for the computational studies.

(4) Choose the value of E_l^{ref} to be used in the constraint equations. This can be determined by simply performing several local constrained optimizations or possibly a short global optimization run with simplified energy models.
(5) Identify initial α values for both the objective and constraint functions.
(6) Set initial best upper bound to an arbitrarily large value.

The *computation* phase is iterative in nature, and depends on the refinement of the original domain through partitioning along the global variables. In each subdomain, upper and lower bounding problems are solved locally and used to develop the sequence of converging upper and lower bounds. The basic steps are as follows:

(1) The original domain is partitioned along one of the global variables.
(2) Lower bounding functions for both the objective and constraints are constructed in both subdomains. A constrained local minimization is performed using the following procedure :

(A) 100 random points are generated and used for evaluation of the lower bounding objective function and constraints. The point with the minimum objective function value is used as a starting point for local minimization using NPSOL.
(B) If the minimum value found is greater than the current best upper bound the subdomain can be fathomed (global minimum is outside region), otherwise the solution is stored.

(3) The upper bounding problems (original constrained formulation) are then solved in both subdomains according to the following procedure :

(A) Set counter, $c = 1$. 100 random points are generated and used for evaluation of the objective function and constraints. The point with the minimum objective function value and smallest violation of the constraints is used as a starting point to perform TAD (1000 high temperature steps followed by 3000 annealing steps) using the simplified target function. The torsion angle bounds of the current subdomain determine the dihedral angle restraint functions. In addition to the NOE derived distance restraints, sterically based distance restraints are added to prevent van der Waals overlaps.

(i) If the $E_l^{\text{distance}} < E_l^{\text{ref}} \ \ \forall l = 1, \ldots, N_{\text{CON}}$, go to step 3.B. Else go to step 3.A.ii.
(ii) Increment counter, $c = c+1$. If $c < 5$, reduce weight of sterically based distance restraints, perform new TAD and go to step 3.A.i. Else go to step 3.B.

(B) Set counter, $c = 1$. Perform local minimization using NPSOL with dihedral angle box constraints to implicitly enforce bounds. The objective func-

tion is a weighted combination of forcefield energy and distance restraint terms :

$$E = E_{\text{ECEPP/3}} + \sum_l W_l E_l^{\text{distance}}. \tag{14}$$

where the weights, W_l, are based on the violation of the distance constraints :

$$W_l = \sqrt{1 + \frac{E_l^{\text{distance}}}{E_l^{\text{ref}}}}. \tag{15}$$

(i) If $E_l^{\text{distance}} < E_l^{\text{ref}} \quad \forall l = 1, \ldots, N_{\text{CON}}$, go to step 3.C. Else go to step 3.B.ii.

(ii) Increment counter, $c = c + 1$. If $c < 5$, increase weight of distance restraint terms, perform TAD (100 high temperature steps followed by 300 annealing steps) and go to step 3.B.i. Else go to step 3.C.

(C) Solve the constrained minimization problem using NPSOL and store all feasible solutions.

(4) The current best upper bound is updated to be the minimum of those thus far stored.

(5) The subdomain with the current minimum value of $L_{\text{forcefield}}$, as given by Equation (12), is selected and partitioned along one of the global variables.

(6) If the best upper and lower bounds are within a defined tolerance the program will terminate, otherwise it will return to Step 2.

5.3. IMPLEMENTATION ISSUES : DISTRIBUTED COMPUTING

The final stage of the approach employs a combination of the deterministically based αBB global optimization algorithm and molecular dynamics in torsion angle space to solve a constrained tertiary structure prediction problem. Since the torsion angle dynamics serves as an initialization strategy in the αBB algorithm, parallelization simply mirrors the overall branch and bound nature of the approach. A characteristic of a branch and bound framework is that as the size of the domain decreases, the quality of the representation improves, which implies that finer initial domains should result in better approximations. This is equivalent to simultaneously exploring multiple domains in order to perform a more efficient search, which coincides with the rationale behind advocating the development of a parallel algorithm.

Distributed frameworks for branch-and-bound algorithms can rely on two basic protocols. The simplest conceptual structure consists of a tree hierarchy in which a master processor directs the overall flow of the algorithm. In this case, global

communication constructions can be maintained in order to control termination and domain processing. The second alternative relies on a ring structure in which all processors act locally and utilize predetermined communication patterns to relay information and detect termination.

Due to the significant computational effort required to initialize and solve the constrained tertiary structure prediction problems for a single node in the branch and bound tree, communication overhead does not substantially affect overall processing time. That is, the time spent in solving the lower and upper bounding problems for each region is long relative to the time required for communication. Therefore, a simple tree hierarchy through a master–slave decomposition approach has been implemented. The implementation requires the creation of only one communication group in which a single master processor maintains the list of lower bounds. The initial domains for the slave nodes are determined by the master through partitioning of the global domain to the appropriate level in the branch-and-bound tree, and these regions are sent to the nodes for further processing. Once the upper and lower bounding problems have been solved, the relevant information is returned to the master, which extracts and sends to the idle node the next region from the lower bound list. The local processing of each domain may also encompass several levels in the branch and bound tree depending on the computational requirements for solving one node in the tree.

Several factors affect the computational requirements for solving this constrained tertiary structure prediction problem. Most notable are the form of the energetic model, the form of the constraint functions and the number of global variables for the system. For a system of approximately 60 residues, the tertiary structure prediction phase, as described above, requires two days of CPU time on a 80 processor distributed computing environment running Linux (16 Pentium-III 450 MHz and 64 Pentium-III 600 MHz processors).

6. Computational Studies

6.1. BOVINE PANCREATIC TRYPSIN INHIBITOR

The approach for tertiary structure prediction was applied to bovine pancreatic trypsin inhibitor (BPTI), a small globular protein found in many tissues throughout the body. BPTI inhibits several of the serine protease proteins such as trypsin, kallikrein, chymotrypsin, and plasmin, and is a member of the pancreatic trypsin inhibitor (kunitz) family, which is a family of serine protease inhibitors. These proteins usually have conserved cysteine residues that participate in the formation of disulfide bonds. In particular, BPTI possesses three disulfide bonds, which are denoted as Cys5–Cys55, Cys14–Cys38, and Cys30–Cys51. The structure of the 58-amino acid residues chain of BPTI has been resolved through several methods, including X-ray crystallography (4PTI) (Deisenhofer and Steigemann, 1975) and a combination of X-ray and neutron diffraction experiments (5PTI) (Wlodawer et

al., 1984). Basic secondary structural features include a N-terminal 3_{10} helix, a C-terminal α-helix and several antiparallel β-strand configurations.

For BPTI tertiary structure prediction the α-helix and β-sheet prediction results in defined dihedral angle domains for 30 of the 58 total residues. The α-helical ϕ-ψ domain was assigned to residues 2–5 and 47–54, while the β-strands between residues 17–23, 29–35 and 44–46 assumed dihedral angle bounds from the extended region of the ϕ-ψ domain space. Lower and upper C^{α}–C^{α} distance restraints were introduced to enforce α-helix hydrogen bonding and β-sheet formation. An additional six upper and lower distances were placed on the S^{γ} atoms between cystine residues to enforce the correct disulfide bridge network. These distance constraints were used to formulate one nonconvex constraint, as represented by the 325 dimensions for the dihedral angle variables of the BPTI system.

Using these restraints, the combined global optimization approach and torsion angle dynamics protocol was applied to the BPTI structure prediction problem. During the course of the global optimization search, the branch and bound tree was formed by partitioning domains belonging to the 52 ϕ and ψ variables of the undefined (loop) residues. Along with the global minimum energy structure, a set of low energy solution structures was identified. To gauge these results, comparisons between the crystallographically derived structure and the predicted structures were based on RMSD (root mean squared deviations) between the C^{α} atoms. A significant sample of low energy structures with C^{α} RMSD (root mean squared deviation) values below 6.0 $\AA$ was identified along with the global minimum energy structure.

The lowest energy structure, with an energy of -428.0 kcal/mol, also provided the best superposition with the crystallographic structure, with a 4.1 $\AA$ RMSD (see Figure 5).

6.2. IMMUNOGLOBULIN BINDING DOMAIN OF PROTEIN G

Protein G is a small globular protein produced by several streptococcal species. The proteins are composed of two or three nearly identical domains of about 55 amino acids each. The system considered here is the immunoglobulin-binding domain from streptococcal protein G, a 56-amino acid polypeptide. The structure contains an efficiently packed hydrophobic core between a four-stranded β-sheet and a four-turn α-helix (Gronenborn et al., 1991) with an overall secondary structure of $\beta\beta\alpha\beta\beta$. The formation of the β-sheet consists of two β-hairpin turns, each connecting antiparallel strands. The first and last strands combine to form the final parallel β-sheet to give the four-stranded configuration. Experimental structures have been determined using both crystallographic (Gallagher et al., 1994) and NMR-derived (Gronenborn et al., 1991) data.

Analysis of the immunoglobulin binding domain of Protein G has also been the focus of theoretical studies on protein folding. In particular, the third and fourth β-strands have been used to model the formation of β-sheet structure through

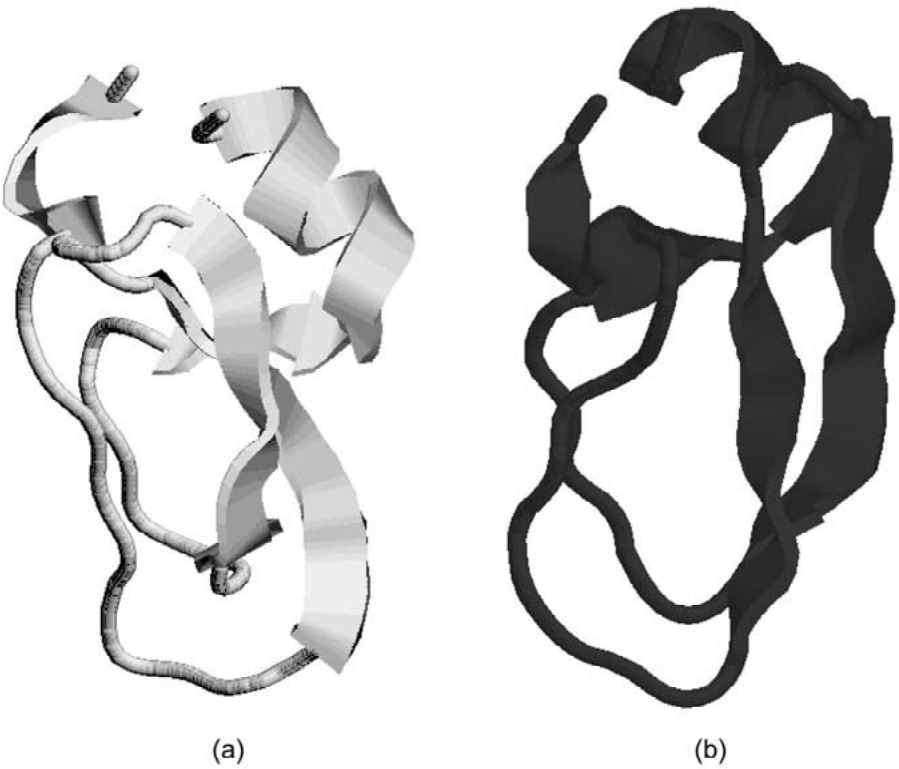

Figure 5. Comparison of predicted lowest energy tertiary structure (in black) of BPTI and experimentally derived structure (in grey). The structures begin with the N-termini at the upper right hand corner of the figure and end with the C-termini at the upper left hand corner of the figure.

hairpin folding. Initial observations included the proposal of a simple statistical mechanical model in which the formation of hydrogen bonds, through a zipper mechanism, drives hairpin folding (Munoz et al., 1997). More recently, simulations have shown that an early step in hairpin folding is the formation of a hydrophobic cluster (Dinner et al., 1999; Pande and Rokhsar, 1999; Bryant et al., 2000).

For the immunoglobulin binding domain of Protein G the α-helix and β-sheet prediction results were used to set dihedral angle domains for 33 of the 56 total residues. The α-helical ϕ–ψ domain was assigned to residues 23–34, while the β-strands between residues 1–7, 16–21, 43–45 and 51–55 were assigned dihedral angle bounds from the extended region of the ϕ-ψ domain space. 22 lower and upper C^{α}–C^{α} distance restraints were introduced to enforce α-helix hydrogen bonding and β-sheet formation. The tertiary structure prediction problem was formulated using one nonlinear constraint for the distance restraints across the total 332 dimensions of the nonconvex dihedral angle hyperspace.

The 23 residues not belonging to helices or strands were treated as loop residues. This provided a set of 56 ϕ and ψ variables upon which the branch and bound tree

was constructed. The bounds of these residues were reduced through the analysis of free energy runs for oligopeptides. The combined global optimization approach identified an ensemble of low energy structures, which were compared to the experimentally derived structure by performing RMSD (root mean squared deviations) calculations between the C^{α} atoms. As with BPTI, a significant sample of low energy structures with RMSD values below 6.0 $\AA$ were discovered. For the immunoglobulin binding domain of Protein G, the structure exhibiting the best RMSD value of 4.2 $\AA$ also provided the lowest energy value.

The superpositioning of the structure exhibiting the lowest RMSD with the experimentally determined structure is shown in Figure 6. The superposition provides an RMSD of 4.2 $\AA$ RMSD, while the structure assumes an energy of -267.0 kcal/mol.

7. Conclusions

The multi-stage ASTRO-FOLD approach provides a general framework for ab-initio prediction of the three-dimensional structure of a protein. The method is grounded on physical insight into the formation of helical and β-structure, and exploits this information in the formulation and solution of the overall structure prediction problem.

In the first stage, detailed free energy calculations including both solvation and ionization effects, are used to identify initiation and termination sites of helices. The second stage, the prediction of β-structure, relies on the principles of a hydrophobically driven collapse to formulate an integer linear optimization problem to identify β-sheet and disulfide bridge connectivity.

The final stages exploit information from the first two stages to develop restraints on the overall tertiary structure prediction problem. These restraints are similar in form to the structure prediction problem using experimental data, and the formulation and solution of this problem borrows from advancements made in this area. Both the novel modeling and search components enhance the applicability and increase the prediction accuracy of the approach over competing solution schemes. The agreement between experimental and predicted structures makes ASTRO-FOLD a very promising method for generic tertiary structure prediction of polypeptides.

Appendix

The TAD algorithm solves for the torsional accelerations, $\ddot{\boldsymbol{\phi}}$:

$$M(\boldsymbol{\phi})\ddot{\boldsymbol{\phi}} + C(\boldsymbol{\phi}, \dot{\boldsymbol{\phi}}) = 0. \tag{16}$$

In this equation M is an $N \times N$ nonlinear mass matrix and C is the N dimensional vector of velocity dependent (Coriolis and other) forces. $\boldsymbol{\phi}$, $\dot{\boldsymbol{\phi}}$ and $\ddot{\boldsymbol{\phi}}$ represent the torsional position, velocities and accelerations, respectively. The ability to calculate the accelerations recursively relies on the chainlike structure of the protein, in

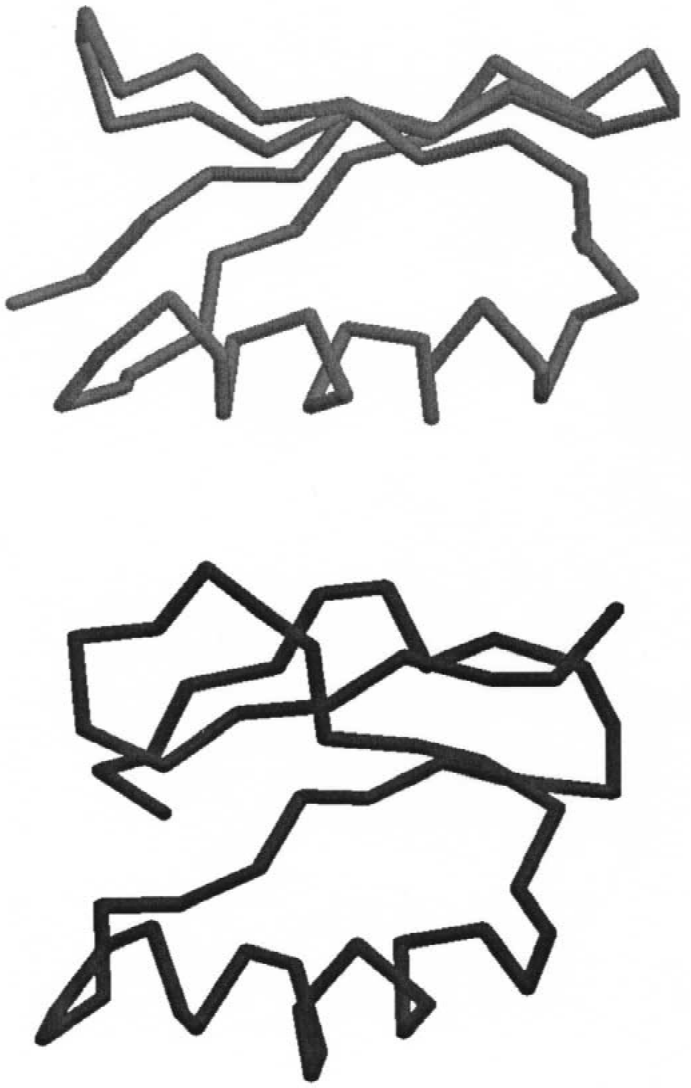

Figure 6. Comparison of predicted (lowest RMSD) tertiary structure (bottom) of immunoglobulin binding domain of Protein G and experimentally derived structure (top).

which each node of the chain represents a rigid body. These rigid bodies consist of one atom or a cluster of atoms whose relative positions are fixed. To simplify the explanation of the algorithm, an unbranched chain will be considered, although the approach can be easily extended to branched systems. For this simple case, the first rigid body, at one end of the chain, defines the base ($k = 0$), while the last rigid body, at the other end of the chain, defines the tip ($k = N$). The rotatable torsion angle between bodies k and $k - 1$ is defined as ϕ_k.

The framework of the algorithm to calculate $\ddot{\boldsymbol{\phi}}$ can be broken down into three steps :

(1) A recursion from the base to the tip is required to calculate the positions, spatial velocities, Coriolis and gyroscopic terms for each of the rigid bodies. To proceed the 6×6 spatial transformation matrix, $\mathbf{s}_k$, between rigid bodies k and $k-1$ must first be defined :

$$\mathbf{s}_k = \begin{bmatrix} I_3 & \tilde{l}(\mathbf{r}_k - \mathbf{r}_{k-1}) \\ 0_3 & I_3 \end{bmatrix}. \tag{17}$$

Here I_3 and O_3 denote the 3×3 dimensional identity and zero matrices, while the $\tilde{l}$ operator refers to the cross product tensor associated with $\mathbf{r}_k - \mathbf{r}_{k-1}$, where $\mathbf{r}_k$ is the position vector that defines the reference frame for rigid body k. The spatial velocity, $\mathbf{V}_k$, can be computed from the following relation :

$$\mathbf{V}_k = \mathbf{s}_k^T \mathbf{V}_{k-1} + \mathbf{H}_k^T \dot{\phi}_k. \tag{18}$$

The spatial velocity is a six-dimensional vector that combines both the three dimensional angular, ω, and linear, $\mathbf{v}$, velocities :

$$\mathbf{V}_k \equiv \begin{pmatrix} \omega_k \\ \mathbf{v}_k \end{pmatrix}. \tag{19}$$

$\mathbf{H}_k$ is also a six dimensional vector with the first three elements corresponding to the unit vector, $\mathbf{e}_k$, in the direction of the bond forming the connection between rigid bodies k and $k-1$:

$$\mathbf{H}_k \equiv \begin{pmatrix} \mathbf{e}_k \\ 0 \end{pmatrix}. \tag{20}$$

The Coriolis and gyroscopic terms, $\mathbf{a}_k$ and $\mathbf{b}_k$, respectively, can then be calculated using the following relationships :

$$\mathbf{a}_k = \begin{pmatrix} 0 \\ \tilde{\omega}_{k-1}[\mathbf{v}_k - \mathbf{v}_{k-1}] \end{pmatrix} + \begin{pmatrix} \tilde{\omega}_k & 0 \\ 0 & \tilde{\omega}_k \end{pmatrix} \mathbf{H}_k^T \dot{\phi}_k \tag{21}$$

$$\mathbf{b}_k = \begin{pmatrix} \tilde{\omega}_k \mathbf{J}_k \tilde{\omega}_k \\ m_k \tilde{\omega}_k \tilde{\omega}_k \mathbf{Y}_k \end{pmatrix}. \tag{22}$$

Both $\mathbf{a}_k$ and $\mathbf{b}_k$ are six dimensional vectors. m_k, $\mathbf{Y}_k$ and $\mathbf{J}_k$ represent the mass, the center of mass vector, and the 3×3 inertia matrix for the rigid body, respectively. Finally, the spatial inertia, $\mathbf{L}_k$, of the rigid body about the reference frame is given by the following 6×6 matrix:

$$\mathbf{L}_k = \begin{pmatrix} \mathbf{J}_k & m_k \tilde{\mathbf{Y}}_k \\ -m_k \tilde{\mathbf{Y}}_k & m_k I_3 \end{pmatrix}. \tag{23}$$

(2) The next step requires a backward recursion from the tip, $k = N$, to the base, $k = 1$. The recursion is used to store a number of auxiliary quantities needed for the final forward recursion to calculate the accelerations. In addition, the gyroscopic terms, $\mathbf{b}_k$, and the spatial inertia terms, $\mathbf{L}_k$, calculated in step 1 can be used to initialize two auxiliary quantities, $\mathbf{z}_k$ and $\mathbf{P}_k$, respectively. Both $\mathbf{P}_k$ and $\mathbf{z}_k$ are updated recursively using the following intermediate terms :

$$D_k = \mathbf{H}_k \mathbf{P}_k \mathbf{H}_k^T \tag{24}$$

$$\mathbf{G}_k = \mathbf{P}_k \mathbf{H}_k^T D_k^{-1} \tag{25}$$

$$\epsilon_k = -\mathbf{H}_k(\mathbf{z}_k + \mathbf{P}_k \mathbf{a}_k) - \nabla E_k. \tag{26}$$

Here D_k and ϵ_k denote scalar quantities, while $\mathbf{G}_k$ is a six dimensional vector. The final equation requires the gradient of the potential function, ∇E_k. The recurrence relationships for $\mathbf{P}_{k-1}$ and $\mathbf{z}_{k-1}$ are given by :

$$\mathbf{P}_{k-1} \leftarrow \mathbf{P}_{k-1} + \mathbf{s}_k(\mathbf{P}_k - \mathbf{G}_k \mathbf{H}_k^T \mathbf{P}_k)\mathbf{s}_k^T \tag{27}$$

$$\mathbf{z}_{k-1} \leftarrow \mathbf{z}_{k-1} + \mathbf{s}_k(\mathbf{z}_k + \mathbf{P}_k \mathbf{a}_k + \mathbf{G}_k \epsilon_k). \tag{28}$$

(3) A final forward recursion from the base to the tip is used to obtain the $\ddot{\boldsymbol{\phi}}$ values. The six dimensional vector α_k is used to store intermediate quantities, with α_k equal to a vector of zeroes for $k = 0$.

$$\alpha_k = \mathbf{s}_k^T \alpha_{k-1} \tag{29}$$

$$\ddot{\phi}_k = \epsilon_k D_k^{-1} - \mathbf{G}_k \alpha_k. \tag{30}$$

The following recursion relation is used to update the values of α_k

$$\alpha_k \leftarrow \alpha_k + \mathbf{H}_k \ddot{\phi}_k + \mathbf{a}_k. \tag{31}$$

For branched molecular structures, each node can potentially spawn more than one child so that both the inward and outward recursions must be modified. In the case of an inward recursion, the results from each of the child nodes must be summed up before moving up one level. In the case of the outward recursion, each of the node branches requires a separate recursion.

The TAD is carried out using simulated annealing, with temperature control provided by coupling to an external bath (Berendsen et al., 1984). This coupling provides a means for forcing or damping the torsional velocities using the following scaling factor at time t:

$$f_T = \sqrt{1 - \frac{1}{\beta} + \frac{T_o}{\beta T(t)}}. \tag{32}$$

In this equation, β is a force constant, while T_o is the bath temperature and $T(t)$ is the actual temperature. The actual temperature is calculated from the kinetic energy, E_{kinetic}, with the following relationship :

$$T(t) = \frac{2E_{\text{kinetic}}(t)}{N_\phi k_{\text{B}}}. \tag{33}$$

where k_{B} is the Boltzmann constant. The value for f_T is used to scale the torsional velocities :

$$\dot{\boldsymbol{\phi}}(t) \leftarrow f_T \dot{\boldsymbol{\phi}}(t). \tag{34}$$

Once torsional velocities have been determined, the accelerations, $\ddot{\boldsymbol{\phi}}$, can be calculated using the recursive algorithm outlined above. As a simple implementation, a basic leap-frog technique is then employed to calculate velocities at the half time-step, which can be used to calculate torsional positions, $\boldsymbol{\phi}$, and new estimated velocities at the full time step.

Acknowledgements

The authors gratefully acknowledge financial support from the National Science Foundation and the National Institutes of Health (R01 GM52032).

References

Adjiman, C. S. and Floudas, C. A. (1996), Rigorous Convex Underestimators for General Twice–Differentiable Problems. *J. Glob. Opt.* 9: 23–40.

Adjiman, C. S., Androulakis, I. P., Maranas, C. D. and Floudas, C. A. (1996), A global optimization method, αBB, for Process Design. *Comput. Chem. Eng.* 20: S419–S424.

Adjiman, C. S., Androulakis, I. P. and Floudas, C. A. (1997), Global optimization of MINLP problems in process synthesis and design, *Comput. Chem. Eng.* 21: S445–S450.

Adjiman, C. S., Dallwig, S., Floudas, C. A. and Neumaier, A. (1998a), A global optimization method for general twice-differentiable NLPs - I. Theoretical advances, *Comput. Chem. Eng.* 22: 1137–1158.

Adjiman, C. S., Androulakis, I. P. and Floudas, C. A. (1998b), A global optimization method for general twice-differentiable NLPs - II. Implementation and computational results, *Comput. Chem. Eng.* 22: 1159–1179.

Androulakis, I. P., Maranas, C. D. and Floudas, C. A. (1995), αBB: A global optimization method for general constrained nonconvex problems, *J. Glob. Opt.* 7, 337–363.

Androulakis, I. P., Maranas, C. D. and Floudas, C. A. (1997), Global minimum potential energy conformation of oligopeptides, *J. Glob. Opt.* 11(1): 1–34.

Baldwin, R. L. (1995), Alpha helix formation by peptides of defined sequence, *Biophys Chem* 55: 127–135.

Baldwin, R. L. and Rose, G. D. (1999a), Is protein folding hierarchic? I. Local structure and peptide folding, *Trends Biochem. Sci.* 24: 26–33.

Baldwin, R. L. and Rose, G. D. (1999b), Is protein folding hierarchic? II. Folding intermediates and transition states, *Trends Biochem. Sci.* 24: 77–83.

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. and Haak, J. R. (1984), Molecular dynamics with coupling to an external bath, *J. Chem. Phys.* 81: 3684–3690.

Brünger, A. (1992), X-PLOR, *Version 3.1 A system for X-ray Crystallography and NMR*. Yale University Press, New Haven, USA.

Bryant, Z., Pande, V. S. and Rokhsar, D. S. (2000), Mechanical unfolding of a beta-hairpin using molecular dynamics, *Biophys J* 78: 584–589.

Caves, L. S. D., Evanseck, J. D. and Karplus, M. (1998), Locally accessible conformations of proteins: Multiple molecular dynamics simulations of cramb in, *Protein Sci.* 7: 649–666.

Chan, C.-K., Hu, Y., Takahashi, S., Rousseau, D. L., Eaton, W. A. and Hofrichter, J. (1997), Submillisecond protein folding kinetics studied by ultrarapid mixing, *Proc. Nath. Acad. Sci., USA* 94: 1779–1784.

Daggett, V., Li, A. J. and Fersht, A. R. (1998), Combined molecular dynamics and Phi-value analysis of structure-reactivity relationships in the transition state and unfolding pathway of barnase: Structural basis of Hammond and anti-Hammon d effects, *J. Am. Chem. Soc.* 120: 12740–12754.

Deisenhofer, J. and Steigemann, W. (1975), Crystallographic refinement of the structure of bovine pancreatic trypsin inhibitor at 1.5 A resolution, *Acta Crystallogr. Sect B* 31: 238–250.

Dill, K. A. (1999), Polymer principles and protein folding, *Prot. Sci.* 8: 1166–1180.

Dinner, A. R., Lazaridis, T. and Karplus, M. (1999), Understanding beta-hairpin formation, *Proc. Nath. Acad. Sci., USA* 96: 9068–9073.

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

Eyrich, V.A., Standley, D. M., Felts, A. K. and Friesner, R. A. (1999), Protein tertiary structure prediction using a branch and bound algorithm, *Proteins* 35: 41.

Floudas, C. A. (1997), Deterministic global optimization in design, control, and computational chemistry. In: Biegler, L., Coleman, T., Conn, A. and Santosa, F. (eds.), *Large Scale Optimization with Applications, Part II: Optimal Design and Control*, Vol. 93. pp. 129–184.

Floudas, C. A. (1995), *Nonlinear and Mixed-Integer Optimization*. New York: Oxford University Press.

Floudas, C. A. (2000), *Deterministic Global Optimization: Theory, Methods and Applications*, Nonconvex Optimization and its Applications. Kluwer Academic Publishers, Dordrecht.

Gallagher, T., Alexander, P., Bryan, P. and Gilliland, G. L. (1994), Two crystal structures of the B1 immunoglobulin-binding domain of Streptococcal protein G and comparison with NMR, *Biochemistry* 33: 4721–4729.

Gill, P. E., Murray, W., Saunders, M. A. and Wright, M. H. (1986) NPSOL 4.0 User's Guide, Systems Optimization Laboratory, Dept. of Operations Research, Stanford University, CA.

Gronenborn, A. M., Filpula, D. R., Essig, N. Z., Achari, A., Whitlow, M., Wingfield, P. T. and Clore, G. M. (1991), A novel highly stable fold of the immunoglobulin binding domain of streptococcal protein G *Science* 253: 657–660.

Güntert, P., Mumenthaler, C. and Wüthrich, K. (1997), Torsion angle dynamics for NMR structure calculation with the new program DYANA, *J. Mol. Biol.* 273: 283–298.

Hamada, D., Segawa, S. and Goto, Y. (1996) Non-native alpha helical intermediate in the refolding of beta-lactoglobulin: A predominantly beta sheet protein, *Nat Struct Biol* 3: 868–873.

Hertz, D., Adjiman, C. S. and Floudas, C. A. (1999), Two results on bounding the roots of interval polynomials, *Comput. Chem. Eng* 23: 1333–1339.

Jain, A., Vaidehi, N. and Rodriguez, G. (1993), A fast recursive algorithm for molecular dynamics simulation, *J. Comp. Phys.* 106: 258–268.

Kirkpatrick, S., C. D. G. Jr., and Vecchi, M. P. (1983), Optimization by simulated annealing, *Science* 220: 671–680.

Klepeis, J. L. and Floudas, C. A. (1999), Free energy calculations for peptides via deterministic global optimization, *J Chem Phys* 110: 7491–7512.

Klepeis, J. L. and Floudas, C. A. (2002a), Ab initio prediction of helical segments in polypeptides, *J. Comp. Chem.* 23: 245–266.

Klepeis, J. L. and Floudas, C. A. (2002b), Ab initio prediction of β-sheets and disulfide bridges in polypeptides, Submitted for publication.

Klepeis, J. L., Androulakis, I. P. Ierapetritou, M. G. and Floudas, C. A. (1998), Predicting solvated peptide conformations via global minimization of energetic atom–to–atom interactions, *Comput. Chem. Eng.* 22: 765–788.

Klepeis, J. L., Floudas, C. A., Morikis, D. and Lambris, J. D. (1999), Predicting peptide structures using NMR data and deterministic global optimization, *J Comp Chem* 20: 1354–1370.

Klepeis, J. L., Pieja, M. T. and Floudas, C. A. (2002), Algorithmic improvements for protein structure prediction via deterministic global optimization, (in preparation).

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

Lim, W. A. and Sauer, R. T. (1991), The role of internal packing interactions in determining the structure and stability of a protein, *J Mol Biol* 219: 359–376.

Maranas, C. D. and Floudas, C. A. (1992), A global optimization approach for Lennard-Jones microclusters, *J. Chem. Phys.* 97(10): 7667–7677.

Maranas, C. D. and Floudas C. A. (1993), Global optimization for molecular conformation problems, *Annals of Operations Research* 42: 85–117.

Maranas, C. D. and Floudas, C. A. (1994a), A deterministic global optimization approach for molecular structure determination, *J. Chem. Phys.* 100(2): 1247–1261.

Maranas, C. D. and Floudas, C. A. (1994b), Global minimum potential energy conformations of small molecules, *J. Glob. Opt.* 4: 135–170.

Maranas, C. D., Androulakis, I. P. and Floudas, C. A., (1996), A deterministic global optimization approach for the protein folding problem, in: *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, Vol. 23. pp. 133–150.

Munoz, V. and Serrano, L. (1994), Elucidating the folding problem of helical peptides using empirical parameters, *Nat Struct Biol* 1: 399–409.

Munoz, V., Thompson, P. A., Hofrichter, J. and Eaton, W. A. (1997), Folding dynamics and mechanism of beta-hairpin formation, *Nature* 390: 196–199.

Némethy, G., Gibson, K. D., Palmer, K. A., Yoon, C. N., Paterlini, G., Zagari, A., Rumsey, S. and Scheraga, H. A. (1992), Energy parameters in polypeptides. 10, *J. Phys. Chem.* 96: 6472.

Pande, V. S. and Rokhsar, D. S. (1999), Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein G, *Proc. Nath. Acad. Sci., USA* 96: 9062–9067.

Rice, L. M. and Brünger, A. T. (1994), Torsion angle dynamics: Reduced variable conformational sampling enhances crystallographic structure refinement, *Proteins* 19: 277–290.

Scheraga, H., 1996, PACK: Programs for packing polypeptide chains, online documentation.

Srinivasan, R. and Rose, G. D. (1999), A physical basis for protein secondary structure, *Proc. Nath. Acad. Sci., USA* 96: 14258–14263.

Standley, D. M., Eyrich, V. A., Felts, A. K., Friesner, R. A. and McDermott, A. E.: (1999), A branch and bound algorithm for protein structure refinement from sparse NMR data sets, *J. Mol. Bio.* 285: 1691–1710.

Wlodawer, A., Walter, J., Huber, R. and Sjolin L. (1984), Structure of bovine trypsin inhibitor. Results of joint neutron and X-ray refinement of crystal form II, *J. Mol. Biol.* 180: 301–329.