Deterministic Global Optimization and Ab-Initio Approaches for the Structure Prediction of Polypeptides, Dynamics of Protein Folding and Protein-Protein Interactions

John L. Klepeis, Heather D. Schafroth,
Karl M. Westerberg and Christodoulos A. Floudas

Department of Chemical Engineering, Princeton University
Princeton, N.J. 08544-5263

Contents

I Introduction

Proteins are some of the most complex and vital molecules in nature. Their complexity arises from the intricate balance of intra- and inter-molecular interactions which define their native three-dimensional structures and biological functionalities. Recent advances in genetic engineering and genome projects have heightened interest in predicting the folding dynamics and equilibrium structures of proteins and protein-protein complexes. This prediction ability is of great theoretical interest, especially in the fields of biophysics and biochemistry. The applications of these predictions promise to be especially valuable. The ability to predict the structure of individual and complexed protein molecules would increase our understanding of disease, aid in the interpretation of genome data, and revolutionize the process of de novo drug design.

Anfinsen's thermodynamic hypothesis [13] suggests that the native structure of a protein system is in a state of thermodynamic equilibrium corresponding to the system with the lowest free energy. Experimental studies have shown that, under native physiological conditions and after denaturation, globular proteins spontaneously refold to their unique, native structure [125]. Understanding the transition of a protein from a disordered state to its native state defines the protein folding problem. A natural extension of the protein folding problem is the related problem of predicting protein-protein interactions, also known as peptide docking. Prediction of protein-protein interactions requires the identification of equilibrium structures for protein-protein complexes. One part of this prediction challenge involves identifying the conformation of the binding sites through which complexed protein interact, and this can be accomplished experimentally or approached as an independent protein folding problem. Another part of the peptide docking prediction challenge involves identifying equilibrium structures for a number of candidate "docking" molecules complexed with a target macromolcule and then quantifying and comparing their relative binding affinities.

The use of computational techniques and simulations in addressing the protein folding and peptide docking problems became possible through the introduction of qualitative and quantitative methods for modeling these systems. The development of realistic energy models also established a link to the field of global optimization, where, based on Anfinsen's hypothesis, the quantity to be optimized is the free energy of the system. Because the number of local minima is vast, the corresponding problem formulation has earned the simple yet suggestive title of the "multiple-minima" problem. The basis for these difficulties is best summarized by Levinthal's paradox [148]. This paradox suggests a contradiction between the almost infinite number of possible stable states that the system may sample and the relatively short time scale required for actual protein folding. Levinthal's observations suggest that the native state is the lowest kinetically accessible free energy minimum, which may be different from the true global minimum. These principles have been used to develop computational techniques for predicting protein folding pathways [21, 50, 44, 147, 238]. Such techniques attempt to map the shape of the energy hypersurface and determine whether this surface "funnels" a protein towards a dominant conformational basin. By invoking the thermodynamic hypothesis, the overall shape of the energy hypersurface is neglected and the problem can be formulated in terms of global minimization, which requires the use of effective global optimization techniques. If this formulation is to reproduce the behavior of realistic systems, the folding of actual proteins should not be kinetically hindered. This has been verified for various systems by performing denaturation–refolding experiments. In addition, by introducing structural characteristics whose formation may act as kinetic barriers, such as the formation of disulfide bonds, the performance of the thermodynamic equilibrium model should be improved.

To better understand the dynamics of protein folding it is also necessary to examine a protein's energy hypersurface. The characterization of the energy surface must include the identification of other stable and metastable configurations. Mathematically, these structures correspond to stationary points of the energy function. In particular, local minima represent stable conformations, while (first order or higher order) saddle points constitute transition states that connect two stable structures. A folding pathway defines the connection between two stable conformations (local minima) through a series of transition states (saddle points). Since the folding pathway may include a number of intermediates, a rigorous description of the energy surface would require the identification of all local minima and saddle points of the energy function.

Based on the complexity of the energy hypersurface, there is an obvious need for the development of efficient global optimization techniques. Although the energy can be expressed analytically, exhaustive searches are possible for only the smallest of systems. These observations, and the importance of the protein folding and peptide docking problems, have propelled the introduction of new global search strategies specifically designed for these problems.

In the sequel, we first outline the basics of the deterministic global optimization approach, αBB, which has been used extensively to study the protein structure prediction, dynamics of protein protien folding and protein docking problems. This is followed by a comprehensive study of ab-initio modeling for structure prediction of single chain polypeptides in Section III. An extensive comparison of energy modeling, including solvation, entropic effects and free energy calculations, is provided for the oligopeptides. The related problem of restrained structure refinement in the presence of sparse experimentally derived restraints is also discussed. Section IV moves beyond the static structure prediction problem towards an understanding of the dynamics of protein folding. An in-depth analysis of the coil to helix transition is provided for the alanine tetrapeptide. This analysis includes the elucidation of folding pathways and the identification of plausible reaction coordinates. Section V addresses the peptide docking problem. First, an approach for the determination of binding site structure is introduced. This is followed by a decomposition-based approach for the prediction of relative binding affinities. Both approaches are applied to peptide docking in HLA molecules.

II Deterministic Global Optimization

A Twice Continuously Differentiable NLPs

The generic optimization problem to be addressed has the following form:

$$\begin{array}{cr} \min\limits_{\mathbf{x}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{g}(\mathbf{x}) \leq 0 \\ & \mathbf{h}(\mathbf{x}) = 0 \\ & \mathbf{x} \in [\mathbf{x}^L, \mathbf{x}^U] \end{array} \tag{1}$$

where $\mathbf{x}$ is a vector of n continuous variables, $f(\mathbf{x})$ is the objective function, $\mathbf{g}(\mathbf{x})$ is a vector of inequality constraints, and $\mathbf{h}(\mathbf{x})$ is a vector of equality constraints. Both the objective function and constraint equations are assumed to be twice continuously differentiable. $\mathbf{x}^L$ and $\mathbf{x}^U$ denote the lower and upper bounds on the $\mathbf{x}$ variables, respectively. The constraints define the feasible region for the problem.

Two main classes of global optimization techniques have been developed to address problem (1), namely, stochastic and deterministic approaches. Stochastic methods, such as those based on genetic algorithms [84] and simulated annealing [126], can be used to treat unconstrained nonconvex problems. However, the stochastic nature of the search strategy invalidates any claims regarding global optimality since it is impossible to obtain valid bounds on the solution of the problem. The addition of nonconvex constraints further complicates these solution schemes. In contrast, deterministic methods rely on a theoretically-based search of the domain space to guarantee the identification of the global optimum solution.

A common characteristic of deterministic global optimization algorithms is the progressive reduction of the domain space until the global solution has been found with arbitrary accuracy. The solution is approached from above and below by generating converging sequences of upper and lower bounds, and the generation of these bounds on the global optimum solution is an essential part of all deterministic global optimization algorithms [71, 103, 104].

The αBB algorithm has been developed to address general twice continuously differentiable models of type 1 [5, 3, 4, 2, 9]. The algorithm is built on a branch-and-bound framework and can handle generic nonconvex optimization problems represented by formulation (1). ϵ-convergence to the global optimum solution is guaranteed when the functions $f(\mathbf{x})$, $\mathbf{g}(\mathbf{x})$ and $\mathbf{h}(\mathbf{x})$ are twice continuously differentiable. The algorithm has been shown to terminate in a finite number of iterations for this broad class of problems [4, 2, 157, 158].

The αBB global optimization approach is based on the convex relaxation of the original nonconvex formulation (1). This requires convex lower bounding of all expressions, and these expressions can be classified as : (i) convex terms; (ii) nonconvex terms of special structure; and (iii) nonconvex terms of general structure. Obviously, convex lower bounding functions are not required for original convex expressions (e.g., linear terms). Certain nonconvex terms, including bilinear, trilinear and univariate concave functions, possess special structure that can be exploited in developing lower bounding functions. All other nonconvex terms can be underestimated using a general expression [9].

When applying the αBB approach to the protein folding problem, formulation (1) involves only nonconvex expressions of general structure. For this reason, the following exposition will briefly cover underestimation for terms of special structure and then focus on the development of a convex lower bounding formulation for global optimization involving generic nonconvex objective and constraint functions.

1 Underestimating terms of special structure

In the case of a bilinear term xy, [6] showed that the tightest convex lower bound over the domain $[x^L, x^U] \times [y^L, y^U]$ is obtained by introducing a new variable w_B which replaces every occurrence of xy in the problem and satisfies the following relationship:

$$w_B = \max\{x^L y + y^L x - x^L y^L;\ x^U y + y^U x - x^U y^U\}. \tag{2}$$

This lower bound can be relaxed and included in the minimization problem by adding two linear inequality constraints,

$$\begin{array}{rcl} w_B & \geq & x^L y + y^L x - x^L y^L, \\ w_B & \geq & x^U y + y^U x - x^U y^U. \end{array} \tag{3}$$

Moreover, an upper bound can be imposed on w to construct a better approximation of the original problem [161]. This is achieved through the addition of two linear constraints:

$$\begin{array}{rcl} w_B & \leq & x^U y + y^L x - x^U y^L, \\ w_B & \leq & x^L y + y^U x - x^L y^U. \end{array} \tag{4}$$

A trilinear term of the form xyz can be underestimated in a similar fashion [159]. A new variable w_T is introduced and bounded by the following eight inequality constraints:

$$\begin{array}{rcl} w_T & \geq & xy^L z^L + x^L y z^L + x^L y^L z - 2x^L y^L z^L, \\ w_T & \geq & xy^U z^U + x^U y z^L + x^U y^L z - x^U y^L z^L - x^U y^U z^U, \\ w_T & \geq & xy^L z^L + x^L y z^U + x^L y^U z - x^L y^U z^U - x^L y^L z^L, \\ w_T & \geq & xy^U z^L + x^U y z^U + x^L y^U z - x^L y^U z^L - x^U y^U z^U, \\ w_T & \geq & xy^L z^U + x^L y z^L + x^U y^L z - x^U y^L z^U - x^L y^L z^L, \\ w_T & \geq & xy^L z^U + x^L y z^U + x^U y^U z - x^L y^L z^U - x^U y^U z^U, \\ w_T & \geq & xy^U z^L + x^U y z^L + x^L y^L z - x^U y^U z^L - x^L y^L z^L, \\ w_T & \geq & xy^U z^U + x^U y z^U + x^U y^U z - 2x^U y^U z^U. \end{array} \tag{5}$$

Fractional terms of the form x/y are underestimated by introducing a new variable w_F and two new constraints [159] which depend on the sign of the bounds on x.

$$\begin{array}{rcl} w_F & \geq & \left\{ \begin{array}{ll} x^L/y + x/y^U - x^L/y^U & \text{if}\, x^L \geq 0 \\ x/y^U - x^L y/y^L y^U + x^L/y^L & \text{if}\, x^L < 0 \end{array} \right. \\ w_F & \geq & \left\{ \begin{array}{ll} x^U/y + x/y^L - x^U/y^L & \text{if}\, x^U \geq 0 \\ x/y^L - x^U y/y^L y^U + x^U/y^U & \text{if}\, x^U < 0 \end{array} \right. \end{array} \tag{6}$$

For fractional trilinear terms, eight new constraints are required [159]. The fractional trilinear term xy/z is replaced by the variable w_{FT} and the constraints for $x^L, y^L, z^L \geq 0$ are given by

$$\begin{aligned}
w_{FT} &\geq xy^L/z^U + x^Ly/z^U + x^Ly^L/z - 2x^Ly^L/z^U, \\
w_{FT} &\geq xy^L/z^U + x^Ly/z^L + x^Ly^U/z - x^Ly^U/z^L - x^Ly^L/z^U, \\
w_{FT} &\geq xy^U/z^L + x^Uy/z^U + x^Uy^L/z - x^Uy^L/z^U - x^Uy^U/z^L, \\
w_{FT} &\geq xy^U/z^U + x^Uy/z^L + x^Ly^U/z - x^Ly^U/z^U - x^Uy^U/z^L, \\
w_{FT} &\geq xy^L/z^U + x^Ly/z^L + x^Uy^L/z - x^Uy^L/z^L - x^Ly^L/z^U, \\
w_{FT} &\geq xy^U/z^U + x^Uy/z^L + x^Ly/z - x^Ly^U/z^U - x^Uy^U/z^L, \\
w_{FT} &\geq xy^L/z^U + x^Ly/z^L + x^Uy^L/z - x^Uy^L/z^L - x^Ly^L/z^U, \\
w_{FT} &\geq xy^U/z^L + x^Uy/z^L + x^Uy^U/z - 2x^Uy^U/z^L.
\end{aligned} \tag{7}$$

Univariate concave functions are trivially underestimated by their linearization at the lower bound of the variable range. Thus the convex envelope of the concave function $ut(x)$ over $[x^L, x^U]$ is the linear function of x:

$$ut(x^L) + \frac{ut(x^U) - ut(x^L)}{x^U - x^L}(x - x^L). \tag{8}$$

The generation of the best convex underestimator for a univariate concave function does not require the introduction of additional variables or constraints.

2 Underestimating general nonconvex terms

A general nonconvex term $f(\mathbf{x})$ belonging to the class of twice continuously differentiable functions can be underestimated over the entire domain $\mathbf{x} \in [\mathbf{x^L}, \mathbf{x^U}]$ by the function $\hat{f}(\mathbf{x})$ defined as

$$\hat{f}(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^{n} \alpha_i (x_i^L - x_i)(x_i^U - x_i) \tag{9}$$

where the α_i's are non negative scalars.

$\hat{f}(\mathbf{x})$ is a guaranteed underestimator of $f(\mathbf{x})$ because the original nonconvex expression is augmented by the addition of separable quadratic functions which are negative over the entire domain $[\mathbf{x}^L, \mathbf{x}^U]$. Furthermore, since the quadratic term is convex, all nonconvexities in the original term $f(\mathbf{x})$ can be overpowered by using sufficiently large values of the α_i parameters.

The convex lower bounding function $\hat{f}(\mathbf{x})$, defined over the rectangular domain of $\mathbf{x^L} \leq \mathbf{x} \leq \mathbf{x^U}$, possesses a number of important properties which guarantee the convergence of the αBB algorithm to the global optimum solution :

(i) $\hat{f}(\mathbf{x})$ is a valid underestimator of $f(\mathbf{x})$. That is:

$$\forall\, \mathbf{x} \in \left[\mathbf{x}^L, \mathbf{x}^U\right] \text{ it can be shown that } \hat{f}(\mathbf{x}) \leq f(\mathbf{x}),$$

(ii) $\hat{f}(\mathbf{x})$ matches $f(\mathbf{x})$ at all corner points.

(iii) $\hat{f}(\mathbf{x})$ is convex in $\mathbf{x} \in \left[\mathbf{x}^L, \mathbf{x}^U\right]$.

(iv) The maximum separation between the nonconvex term of generic structure, $f(\mathbf{x})$, and its convex relaxation, $\hat{f}(\mathbf{x})$, is bounded and also proportional to the positive α parameters and to the square of the diagonal of the current box constraints :

$$\max_{\mathbf{x}^L \leq \mathbf{x} \leq \mathbf{x}^U} \left[f(\mathbf{x}) - \hat{f}(\mathbf{x}) \right] = \frac{1}{4} \sum_i^n \alpha_i \left(x_i^U - x_i^L \right)^2 . \tag{10}$$

(v) The underestimators constructed over supersets of the current set are always less tight than the underestimator constructed over the current box constraints for every point within the current box constraints.

The key development in the convex lower bounding formulation is the definition of the α parameters. Specifically, the magnitude of the α parameters may be related to the minimum eigenvalue of the Hessian matrix of the nonconvex term $f(\mathbf{x})$.

$$\alpha \geq \max \left\{ 0, -\frac{1}{2} \min_{i, \mathbf{x^L} \leq \mathbf{x} \leq \mathbf{x^U}} \lambda_i(\mathbf{x}) \right\} \tag{11}$$

where $\lambda(\mathbf{x})$ represent the eigenvalues of the Hessian matrix $(H_f(\mathbf{x}))$ for the nonconvex term. An explicit minimization problem can be written to find the minimum eigenvalue (λ_{min}):

$$\begin{array}{ll} \min\limits_{\mathbf{x}, \lambda} & \lambda \\ s.t. & \det \left(H_f(\mathbf{x}) - \lambda I \right) = 0 \\ & \mathbf{x} \in [\mathbf{x}^L, \mathbf{x}^U] \end{array}$$

The solution of this problem is a non-trivial matter for arbitrary nonconvex functions.

One method for the rigorous determination of α parameters for general twice differentiable problems involves interval analysis of Hessian matrices to calculate bounds on the minimum eigenvalue [3, 5]. The difficulties arising from the presence of the variables in the convexity condition can be alleviated through the transformation of the exact $\mathbf{x}$-dependent Hessian matrix to an interval matrix $[H_f]$ such that $H_f(\mathbf{x}) \subseteq [H_f]$, $\forall\, \mathbf{x} \in [\mathbf{x}^L, \mathbf{x}^U]$. The elements of the original Hessian matrix are treated as independent when calculating their natural interval extensions [195, 182]. The interval Hessian matrix family $[H_f]$ is then used to formulate a theorem in which the α calculation problem is relaxed [3]. In other words, a valid lower bound on the minimum eigenvalue can be used to calculate rigorous α values :

$$\alpha \geq \left\{ 0, -\frac{1}{2} \lambda_{min} \left([H_f] \right) \right\} \tag{12}$$

where $\lambda_{min}([H_f])$ is the minimum eigenvalue of the interval matrix family $[H_f]$.

An $\mathcal{O}(n^2)$ method to calculate these α values is the straightforward extension of Gerschgorin's theorem [77] to interval matrices. For a real matrix $A = (a_{ij})$, the well-known theorem states that the eigenvalues are bounded below by λ_{min} such that

$$\lambda_{min} = \min_i \left(a_{ii} - \sum_{j \neq i} |a_{ij}| \right) . \tag{13}$$

For an interval matrix $[A] = ([\underline{a}_{ij}, \overline{a}_{ij}])$, a lower bound on the minimum eigenvalue is given by

$$\lambda_{min} \geq \min_i \left[\underline{a}_{ii} - \sum_{j \neq i} \max \left(|\underline{a}_{ij}|, |\overline{a}_{ij}| \right) \right].$$

This procedure provides a single α value which is valid for all variables.

Non-uniform diagonal shift matrices can be used to calculate a different α value for each variable in order to construct an underestimator of the form shown in Equation (9). The non-zero elements of the diagonal shift matrix can no longer be related to the minimum eigenvalue of the interval Hessian matrix $[H_f]$. If all elements of the scaling vector are set to 1, the equation for the α_i values becomes :

$$\alpha_i = \max \left\{ 0, -\frac{1}{2} \left(\underline{a}_{ii} - \sum_{j \neq i} |a|_{ij} \right) \right\},$$

However, the choice of scaling is arbitrary, and different α_i parameters can be estimated through various scaling techniques.

3 Convexification of feasible region

To obtain a valid lower bound on the global solution of the nonconvex problem, the lower bounding problem generated in each domain must have a unique solution. This implies that the formulation include only convex inequality constraints, linear equality constraints and an increased feasible region relative to that of the original nonconvex problem. The left-hand side of any nonconvex inequality constraint, $g(\mathbf{x}) \leq 0$, in the original problem can simply be replaced by its convex underestimator $\hat{g}(\mathbf{x})$, constructed according to Equation (9), to yield the relaxed convex inequality $\hat{g}(\mathbf{x}) \leq \mathbf{0}$.

For an equality constraint containing general nonconvex terms, the equation obtained by simple substitution of the appropriate underestimators is also nonlinear. Therefore, the original equality $h(\mathbf{x}) = 0$ must be rewritten as two inequalities of opposite signs,

$$\begin{array}{rrl} h^+(\mathbf{x}) = & h(\mathbf{x}) & \leq 0 \\ h^-(\mathbf{x}) = & -h(\mathbf{x}) & \leq 0. \end{array} \tag{14}$$

These two inequalities must then be underestimated independently to give $\hat{h}^+(\mathbf{x})$ and $\hat{h}^-(\mathbf{x})$.

4 Convex lower bounding problem formulation

Summarizing the concepts introduced so far, a convex relaxation for any nonconvex problem of type (1) belonging to the broad class of twice continuously differentiable continuous NLPs can be constructed as

$$\begin{array}{ll} \min\limits_{\mathbf{x}} & \hat{f}(\mathbf{x}) \\ \text{subject to} & \hat{\mathbf{g}}(\mathbf{x}) \leq 0 \\ & \hat{\mathbf{h}}^+(\mathbf{x}) \leq 0 \\ & \hat{\mathbf{h}}^-(\mathbf{x}) \leq 0 \\ & \mathbf{x} \in [\mathbf{x}^L, \mathbf{x}^U] \end{array} \tag{15}$$

where ˆ denotes the convex underestimator of the specified function over the domain $\mathbf{x} \in [\mathbf{x}^L, \mathbf{x}^U]$. Since the inclusion of convex terms and nonconvex terms of special structure has been neglected, these functions involve only α type underestimating expressions. These underestimators are functions of the size of the domain under consideration, and because the αBB algorithm follows a branch-and-bound approach, this domain is systematically reduced at each new node of the tree. Tighter lower bounding functions can therefore be generated by updating the underestimating equations. The lower bounds on the problem form a non-decreasing sequence, and the underestimating strategy is therefore consistent, as required for convergence.

5 Variable bound updates

The quality of the convex lower bounding problem can also be improved by ensuring that the variable bounds are as tight as possible. These variable bound updates can either be performed at the onset of an αBB run or at each iteration.

In both cases, the same procedure is followed in order to construct the bound update problem. Given a solution domain, the convex underestimator for every constraint in the original problem is formulated. The bound problem for variable x_i is then expressed as

$$x_i^{L,NEW} / x_i^{U,NEW} = \left\{ \begin{array}{ll} \min_{\mathbf{x}} / \max_{\mathbf{x}} & x_i \\ \text{subject to} & \hat{\mathbf{g}}(\mathbf{x}) \leq 0 \\ & \mathbf{x}^L \leq \mathbf{x} \leq \mathbf{x}^U \end{array} \right. \tag{16}$$

where $\hat{\mathbf{g}}(\mathbf{x})$ are the convex underestimators of the constraints, and the bounds on the variables, $\mathbf{x}^L$ and $\mathbf{x}^U$ are the best calculated bounds. Thus, once a new lower bound $x_i^{L,NEW}$ on x_i has been computed via a minimization, this value is used in the formulation of the maximization problem for the generation of an upper bound $x_i^{U,NEW}$.

Because of the computational expense incurred by an update of the bounds on all variables, it is often desirable to define a smaller subset of the variables on which this operation is to be performed. The criterion devised for the selection of the branching variables can be used in this instance, since it provides a measure of the sensitivity of the problem to each variable.

6 The αBB algorithm

The global optimization method αBB deterministically locates the global minimum solution of (1) based on the refinement of converging lower and upper bounds. The lower bounds are obtained by the solution of (15), which is formulated as a convex programming problem. Upper bounds are based on the solution of (1) using local minimization techniques.

As previously mentioned, the maximum separation between the generic nonconvex terms and their respective convex lower bounding representations is proportional to the square of the diagonal of the current rectangular partition. As the size of the rectangular domains approach zero, this separation also become infinitesimally small. That is, as the current box constraints $[\mathbf{x}^L, \mathbf{x}^U]$ collapse to a point, the maximum separation between the original objective function of (1) and its convex relaxation in (15) becomes zero. This implies that for the positive numbers ϵ and $\mathbf{x}$

there always exists another positive number δ which, by reducing the rectangular region $[\mathbf{x}^L, \mathbf{x}^U]$ around $\mathbf{x}$ so that $\|\mathbf{x}^U - \mathbf{x}^L\| \leq \delta$, cause the difference between the feasible region of the original problem (1) and its convex relaxation (15) to become less than ϵ. Therefore, any feasible point $\mathbf{x}$ of problem (15), including the global minimum solution, becomes at least ϵ–feasible for problem (1) by sufficiently tightening the bounds on $\mathbf{x}$ around this point.

Once the solutions for the upper and lower bounding problems have been established, the next step is to modify these problems for the next iteration. This accomplished by successively partitioning the initial rectangular region into smaller subregions. The number of variables along which subdivision is required is equal to the number of variables $\mathbf{x}$ participating in at least one nonconvex term of the (1) formulation. The default partitioning strategy used in the algorithm involves successive subdivision of the original rectangle into two sub-rectangles by halving on the midpoint of the longest side of the initial rectangle (bisection). Therefore, at each iteration a lower bound of the objective function (1) is simply the minimum over all the minima of problem (15) in each sub-rectangle of the initial rectangle. In order to ensure lower bound improvement, the sub-rectangle to be bisected is chosen by selecting the sub-rectangle which contains the infimum of the minima of (15) over all the sub-rectangles. This procedure guarantees a non–decreasing sequence for the lower bound. A non–increasing sequence for the upper bound is found by solving the original nonconvex problem (1) locally and selecting it to be the minimum over all the previously recorded upper bounds. Obviously, if the single minimum of (15) for any sub–rectangle is greater than the current upper bound, this sub–rectangle can be discarded because the global minimum cannot be within this subdomain (fathoming step).

Because the maximum separation between the nonconvex terms and their respective convex lower bounding functions is both a bounded and a continuous function of the size of rectangular domain, arbitrarily small feasibility and convergence tolerance limits are attained for a finite sized partition element.

The basic steps of the αBB global optimization algorithm are as follows:

(1) *Initialization*

A convergence tolerance, ϵ_c, and a feasibility tolerance, ϵ_f, are selected and the iteration counter, I, is set to one. The current variable bounds $[\mathbf{x}_I^L, \mathbf{x}_I^U]$ for the first iteration are set equal to the global ones $[\mathbf{x}_0^L, \mathbf{x}_0^U]$. Lower and upper bounds $[f^L, f^U]$ on the global minimum of (1) are initialized and an initial current point is selected from the domain.

(2) *Local solution of nonconvex problem*

The nonconvex optimization problem (1) is solved locally within the current variable bounds $[\mathbf{x}_I^L, \mathbf{x}_I^U]$. If the solution is ϵ_f–feasible, the upper bound f^U is updated as follows :

$$f^U \quad = \quad \min\left(f^U, f_I^U\right)$$

where f_I^U is the objective function value for the current ϵ_f–feasible solution.

(3) *Partitioning of current rectangle*

The current rectangle, $[\mathbf{x}_I^L, \mathbf{x}_I^U]$, is bisected into two sub-rectangles ($r = a, b$) for the variable

(l) with the longest side of the initial rectangle :

$$l_I = \arg\max_i \left(x_{i,I}^U - x_{i,I}^L \right)$$

(4) *Solution of underestimating problems*
The parameters $\alpha_{i,I,r}$ are updated for both rectangles ($r = a, b$). The convex optimization problem (15) is solved inside both sub-rectangles ($r = a, b$) using a nonlinear solver (e.g., MINOS5.4 [179], NPSOL [79]). If a solution $f_{I,r}^L$ is less than the current upper bound, f^U, then it is stored.

(5) *Update of lower bound*
The iteration counter is increased by one, and the lower bound, f^L, is updated to be the minimum solution over the stored solutions from previous iterations. The selected region is erased from the stored set.

$$f^L = \min_{I',r} f_{I',r}^L, \quad r = a, b, \quad I' = 1, \ldots, I-1.$$

(6) *Update bounds*
The bounds of the current rectangle are updated to those of the sub-rectangle containing the previously found solution (f^L).

(7) *Check for convergence*
If $\left(f^U - f^L \right) > \epsilon_c$, then return to STEP 2. Otherwise, ϵ_c–convergence has been reached, and the global minimum solution corresponds to point providing f^U.

Figure 1 diagrams an unconstrained one-dimensional example of the approach. The mathematical proof that the αBB global optimization algorithm converges to the global optimum solution is presented in [157]. In addition to computational chemistry related problems, the αBB approach has been applied to a variety of constrained optimization problems [4, 2, 3, 9].

B Enclosure of All Solutions

The αBB algorithm discussed in the previous section was originally designed to solve global optimization problems. However, this algorithm has also proven to be effective in the solution of non-linearly constrained systems of algebraic equations [159], provided only that the constraints are twice continuously differentiable. The key idea is to reformulate the algebraic system of equations as a global optimization problem that exhibits multiple global solutions and then use the αBB approach as a basis for the enclosure of all solutions. In the following sections, we discuss the enclosure of all solutions.

1 Problem Formulation

In general, a non-linearly constrained system of algebraic equations can be expressed in the form

$$f_i(\mathbf{x}) = 0 , \qquad i = 1, \ldots, N_f$$

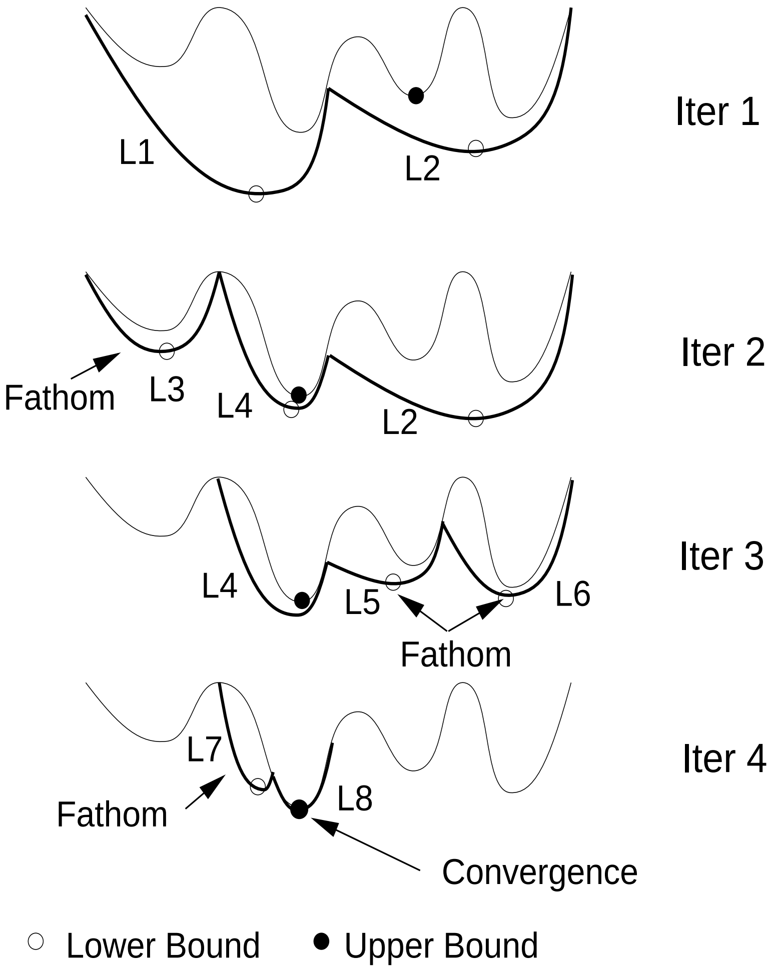

Figure 1: One-dimensional illustrative example of the αBB approach. In iteration 1 the overall domain is bisected, the two convex lower bounding functions are created and their unique minima ($L1$ and $L2$) are identified. An upper bound is also identified. Since $L1$ is less than $L2$, the region containing $L1$ is further bisected in iteration 2, while the other region is stored. The minimum of one region ($L3$) is greater than the new upper bound, so this region can be fathomed. The other region is stored. In iteration 3 the region with the next lowest lower bound ($L2$) is bisected and since both new lower bound minima ($L5$ and $L6$) are greater than the current best upper bound, the entire region is fathomed. Finally, by iteration 4, the region containing $L4$ is bisected which results in a region that can be fathomed (containing $L7$) and a convex region whose minimum ($L8$) equals the current upper bound and is the global minimum.

$$
\begin{aligned}
g_j(\mathbf{x}) &\leq 0 , \qquad j = 1, \ldots, N_g \\
\mathbf{x}^L \leq \mathbf{x} &\leq \mathbf{x}^U
\end{aligned} \tag{17}
$$

where $f_i(\mathbf{x})$ represent the equality constraints (N_f is the number of such constraints) and $g_j(\mathbf{x})$ represent the inequality constraints (N_g is the number of such constraints).

In order to apply the αBB algorithm to (17), we must reformulate it as a global optimization problem. This is accomplished by introducing a slack variable s and minimizing its value over an augmented variable set $(\mathbf{x}, s)$ subject to a set of relaxed constraints:

$$
\begin{aligned}
&\min_{\mathbf{x},s} s \\
\text{subject to} \quad f_i(\mathbf{x}) - s &\leq 0 , \qquad i = 1, \ldots, N_f \\
-f_i(\mathbf{x}) - s &\leq 0 , \qquad i = 1, \ldots, N_f \\
g_j(\mathbf{x}) &\leq 0 , \qquad j = 1, \ldots, N_g \\
\mathbf{x}^L \leq \mathbf{x} &\leq \mathbf{x}^U
\end{aligned} \tag{18}
$$

In comparing the two formulations, the following two facts are self-evident:

- If $s < 0$, the constraints in (18) are infeasible.
- If $s = 0$, the constraints in (18) reduce to the original problem (17).

It follows that $s = 0$ is the *global minimum* of (18) — provided that (17) has solutions — and that there is a one-to-one correspondence between global minima $(\mathbf{x}^*, s^*)$ of (18) and solutions $\mathbf{x}^*$ of the original problem (17). Therefore, the problem of finding all solutions to (17) can be reformulated as the problem of finding all global minima of (18).

In the next section, we will explain how the αBB global optimization algorithm can be used to enclose all global minima of (18), and hence, all solutions to (17).

2 Framework for Enclosing All Solutions

In this section, we describe the αBB global optimization algorithm as it is applied to the general problem of determining all solutions to a system of algebraic constraints (17). This adaptation is based on the correspondence between solutions of (17) and global minima of (18) with $s = 0$. Since the αBB algorithm can be applied to any problem involving constraints which are twice continuously differentiable (C^2), the only necessary assumptions we need to make are that $f_i(\mathbf{x})$ and $g_j(\mathbf{x})$ are C^2 functions for $i = 1, \ldots, N_f$ and $j = 1, \ldots, N_g$, respectively.

The algorithm proceeds by exploring the configuration space for solutions to (17). We begin with the full region $\mathbf{x} \in [\mathbf{x}^L, \mathbf{x}^U]$ and subdivide regions into smaller regions. Each region is tested before it is divided to see if a solution to (17) can possibly exist there. This is accomplished by finding a lower bound of the global minimum of (18) over the region in question. If the lower bound is positive, then $s = 0$ cannot lead to a feasible point of (18), and hence no solution to (17) can exist in the given region. The region will be fathomed (i.e., eliminated from further consideration). On the other hand, if the lower bound is negative or zero, there may or may not be a solution to (17) in that region. In this case, further subdivision and testing will be necessary. If the region size becomes small enough and the region is still active (i.e., its lower bound is negative or zero), then a solution to (17) is obtained within that region by a local search. The algorithm terminates when all regions have been fully processed.

Note that upper bounds of the global minimum need not be determined. Since we are assuming that the global minimum of (18) is zero, we can set the upper bound to this value from the start, and thus avoid the effort of solving an upper bounding problem.

Lower bounds of the global minimum of (18) are determined by solving the lower bounding problem over the given region:

$$\begin{aligned}
&\min_{\mathbf{x},s} s \\
\text{subject to} \quad & \hat{f}_i^{+}(\mathbf{x}) - s \le 0\,, \qquad i = 1,\ldots,N_f \\
& \hat{f}_i^{-}(\mathbf{x}) - s \le 0\,, \qquad i = 1,\ldots,N_f \\
& \hat{g}_j(\mathbf{x}) \le 0\,, \qquad j = 1,\ldots,N_g \\
& \mathbf{x}^L \le \mathbf{x} \le \mathbf{x}^U
\end{aligned} \tag{19}$$

where $\hat{f}_i^{+}(\mathbf{x})$, $\hat{f}_i^{-}(\mathbf{x})$, and $\hat{g}_j(\mathbf{x})$ are convex functions which underestimate $f_i(\mathbf{x})$, $-f_i(\mathbf{x})$, and $g_j(\mathbf{x})$, respectively. Since the constraints are all convex functions, any local optimization package should be able to locate its global minimum. Furthermore, every feasible point of (18) is also a feasible point of (19) since these functions are underestimators of the original functions. It follows that the global minimum of (19) is a valid lower bound of the global minimum of (18).

The crux of the αBB algorithm is finding valid convex underestimators, $\hat{f}_i^{\pm}(\mathbf{x})$ and $\hat{g}_j(\mathbf{x})$, for the functions $\pm f_i(\mathbf{x})$ and $g_j(\mathbf{x})$, respectively, over a given region. An important consideration is that the convex underestimators be as tight (i.e., close in value to the original constraint functions) as is reasonably possible, as tighter underestimators lead to better lower bound estimates. It is important to be able to fathom regions as quickly as possible if they do not contain any solutions to (17). However, this cannot always be done: it frequently occurs that a region contains no solution to (17) (i.e., the global minimum of (18) over that region is positive), but the lower bound obtained from (19) for that region is negative. Such regions obviously must be explored further, until positive lower bounds are obtained. A better lower bound estimate can lead to significant improvement in the efficiency of the algorithm.

When applying this algorithm to the problem of finding all stationary points of a potential energy surface, the constraint functions, $\pm f_i(\mathbf{x})$ and $g_i(\mathbf{x})$, are general nonconvex functions. Whenever these constraint functions are C^2, they can be underestimated using the α underestimation described in Section II.A.2. In this context, the underestimators take the form

$$\begin{aligned}
\hat{f}_i^{+}(\mathbf{x}) &= f_i(\mathbf{x}) - \alpha_i^{f,+} \sum_k (x_k^U - x_k)(x_k - x_k^L) \\
\hat{f}_i^{-}(\mathbf{x}) &= -f_i(\mathbf{x}) - \alpha_i^{f,-} \sum_k (x_k^U - x_k)(x_k - x_k^L) \\
\hat{g}_j(\mathbf{x}) &= g_j(\mathbf{x}) - \alpha_j^{g} \sum_k (x_k^U - x_k)(x_k - x_k^L)
\end{aligned} \tag{20}$$

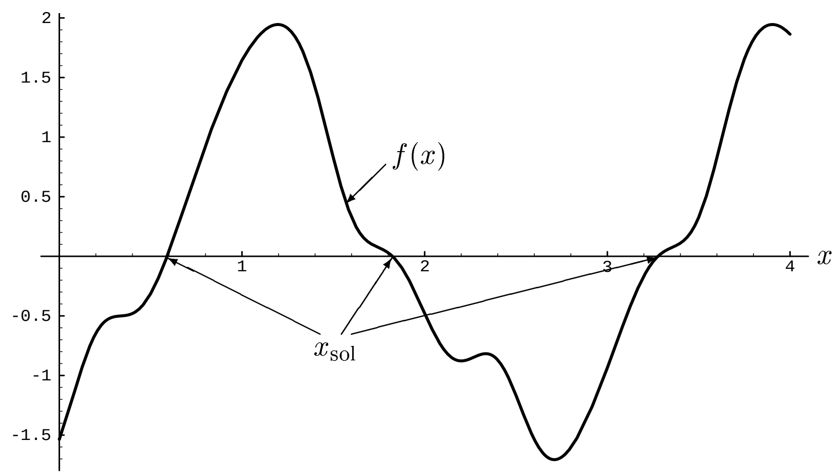

Figure 2: Plot of $f(x)$ for $x \in [0, 4]$.

where the α parameters satisfy the convexity conditions

$$\begin{aligned}
\alpha_i^{f,+} &\geq -\frac{1}{2} \min_{\mathbf{x} \in [\mathbf{x}^L, \mathbf{x}^U]} \{\lambda_k(H_{f_i}(\mathbf{x})), 0\} \\
\alpha_i^{f,-} &\geq +\frac{1}{2} \max_{\mathbf{x} \in [\mathbf{x}^L, \mathbf{x}^U]} \{\lambda_k(H_{f_i}(\mathbf{x})), 0\} \\
\alpha_j^{g} &\geq -\frac{1}{2} \min_{\mathbf{x} \in [\mathbf{x}^L, \mathbf{x}^U]} \{\lambda_k(H_{g_j}(\mathbf{x})), 0\}
\end{aligned} \tag{21}$$

The discussion in Section II.A.2 applies equally well in this situation.

3 Geometrical Interpretation

In this section, we give a geometric illustration of how the αBB algorithm works by showing how it would locate all of the solutions of a single equation $f(x) = 0$ over the interval $x \in [0, 4]$. The function we use for our illustration is

$$f(x) = -2\cos\frac{\pi}{3}(x + 0.05) + e^{-20(x-0.2)^2} - e^{-20(x-1.6)^2} + e^{-20(x-2.4)^2} - e^{-20(x-3.5)^2}$$

A graph of this function is given in Figure 2. There are three solutions to $f(x) = 0$ in this interval. They are

$$x_{\text{sol}} \in \{0.59014, 1.82399, 3.27691\}$$

The corresponding global optimization problem is obtained by introducing a slack variable s and minimizing s subject to the constraints

$$f(x) - s \leq 0 \leq f(x) + s$$

The feasibility region for fixed s is determined by intersecting the region of space between $f(x) - s$ and $f(x) + s$ with the x-axis. This procedure is shown graphically in Figure 3. For $s > 0$, the

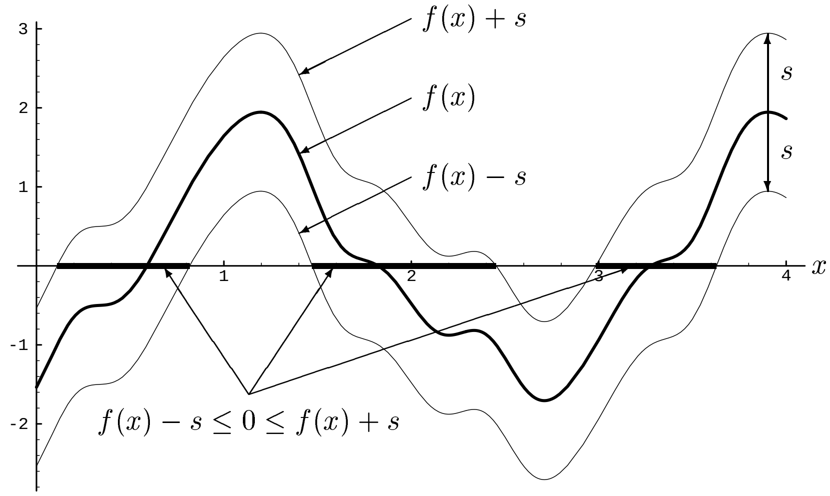

Figure 3: $f(x)$ is shifted by a positive slack variable $s = 1$. Note that the feasibility region of $f(x) - s \leq 0 \leq f(x) + s$ forms intervals around the solutions to $f(x) = 0$.

feasibility region forms intervals around the actual solutions to $f(x) = 0$. Minimizing s subject to the constraints above has the effect of pushing the two graphs together until they both meet at $f(x)$ (at $s = 0$). At $s = 0$, the feasibility region reduces to the solution set for $f(x) = 0$ (each interval reduces to a point). For $s < 0$, the graphs cross and the feasibility region is empty. $s = 0$ is clearly the global minimum whenever $f(x) = 0$ has solutions.

In order to set up the lower bounding problem, we need to find convex underestimators for $\pm f(x)$ for each interval under consideration. We begin with the complete interval $[0, 4]$. The function $f(x)$, and a valid set of convex underestimators $\hat{f}^{\pm}_{[0,4]}(x)$ are plotted in Figure 4 The convex underestimators $\hat{f}^{\pm}_{[0,4]}(x)$ essentially envelop the graph of $f(x)$ in a convex region. This convex region contains all the points $\hat{f}^{+}(x) \leq y \leq -\hat{f}^{-}(x)$, and its intersection with the x-axis is given by $\hat{f}^{+}(x) \leq 0 \leq -\hat{f}^{-}(x)$. All solutions to $f(x) = 0$ in the region $x \in [0, 4]$ must lie in this intersection region because $\hat{f}^{+}(x)$ and $-\hat{f}^{-}(x)$ surround the function $f(x)$ (see Figure 4). If this region had been empty, then no solution to $f(x) = 0$ could possibility exist in the interval $[0, 4]$. This is not the case, but see later on when we discuss the interval $[2, 3]$.

Determining whether or not the feasibility region of $\hat{f}^{+}(x) \leq 0 \leq -\hat{f}^{-}(x)$ is empty involves introducing a slack variable and minimizing it subject to

$$\hat{f}^{+}(x) - s \leq 0 \leq -\hat{f}^{-}(x) + s \tag{22}$$

This is the lower bounding problem. For $s = 0$, (22) reduces to $\hat{f}^{+}(x) \leq 0 \leq -\hat{f}^{-}(x)$. For $s \neq 0$, the feasibility region of (22) is determined by shifting the enveloping functions $\hat{f}^{+}(x)$ and $-\hat{f}^{-}(x)$ by an amount s — away from each other if $s > 0$, and towards each other if $s < 0$ (see Figure 5). Graphically, minimizing s subject to (22) involves expanding or shrinking the region between the underestimators by adjusting s until the region between $\hat{f}^{+}(x) - s$ and $-\hat{f}^{-}(x) + s$ intersects the x-axis at a single point. For the interval $[0, 4]$, this requires moving $\pm\hat{f}^{\pm}(x)$ *towards* each other,

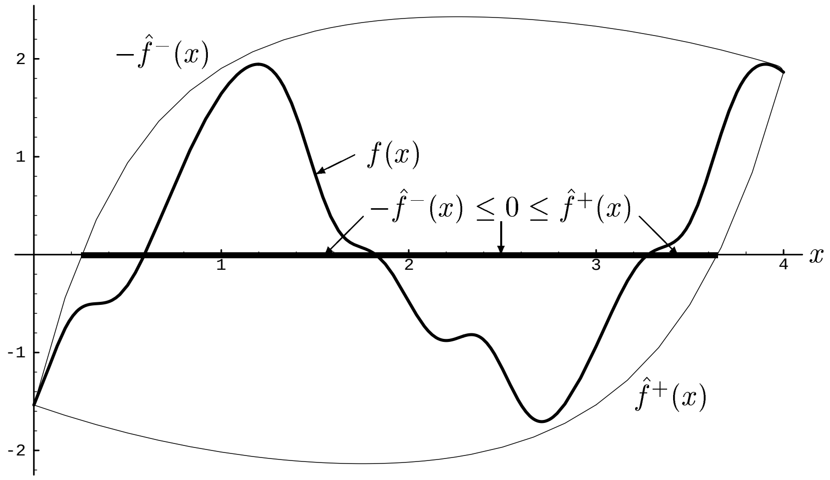

Figure 4: The functions $\hat{f}^{\pm}(x)$ are convex underestimators of $\pm f(x)$ over the interval $[0, 4]$. Note how $\hat{f}^{+}(x)$ and $-\hat{f}^{-}(x)$ form a convex envelope around $f(x)$.

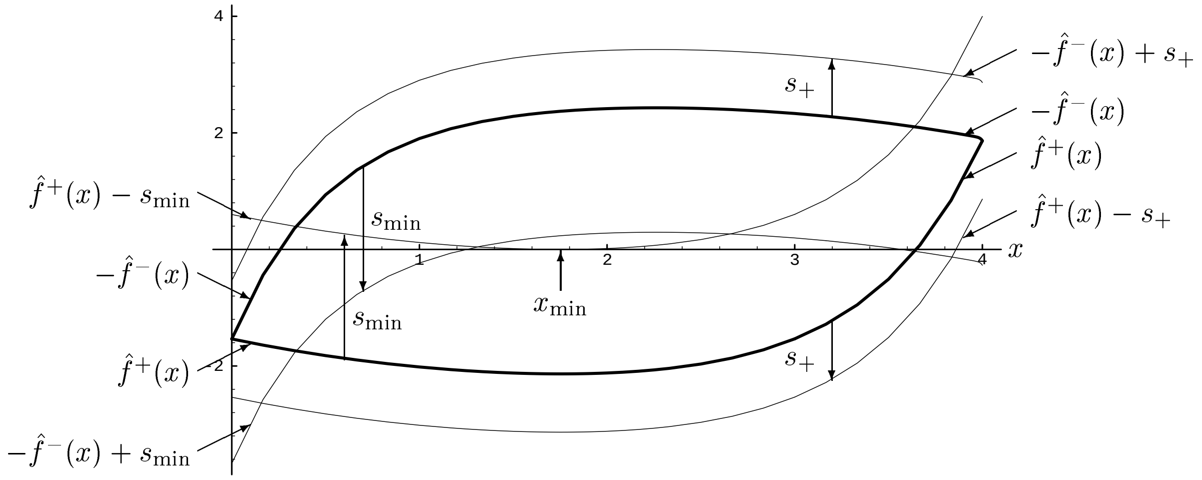

Figure 5: During the solution to the lower bounding problem, the convex underestimators $\hat{f}^{\pm}(x)$ are shifted by a slack variable. Two different shifts are shown above, one positive $s_{+} = 1$, and the other negative $s_{\min} = -2.135$. $s_{\min}$ represents the global minimum to the lower bounding problem: the feasibility region of the lower bounding problem is reduced to a single point $x_{\min} = 1.754$, shown above.

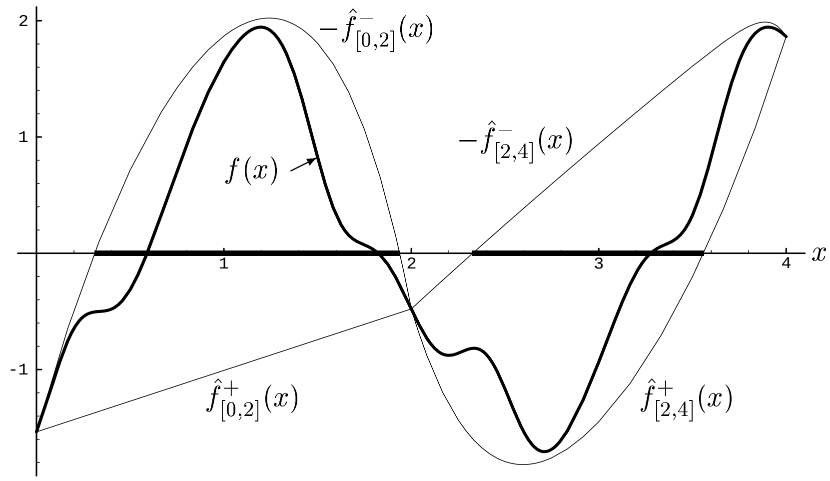

Figure 6: The interval $[0,4]$ has been subdivided into $[0,2]$ and $[2,4]$. The convex underestimators $\hat{f}^{\pm}(x)$ for each subinterval, shown above, form a convex envelope around $f(x)$. As the intervals get smaller, the envelope gets tighter.

implying $s_{\min} < 0$. The fact that $s_{\min} < 0$ indicates that there might be solutions to $f(x) = 0$ in this interval: we will be forced to explore this region further. Note that the lower bounding problem is a *convex problem*, and so any local optimization package should reach this unique global minimum.

We therefore subdivide the interval $[0,4]$ into two subintervals, $[0,2]$ and $[2,4]$, and explore each interval for solutions just as we did for $[0,4]$. The convex underestimators for each interval, $\hat{f}^{\pm}_{[0,2]}(x)$ and $\hat{f}^{\pm}_{[2,4]}(x)$, are shown in Figure 6. Note that each pair of underestimators envelops the corresponding portion of the function $f(x)$, and that the underestimators have improved: they are closer to the function $f(x)$. This will continue to happen as the intervals become narrower.

Again, the question we ask in each interval is: can a solution to $f(x) = 0$ exist there? The question is answered by solving the lower bounding problem. In both cases, the region $\hat{f}^{+}(x) \leq 0 \leq -\hat{f}^{-}(x)$ does intersect the x-axis (see Figure 6), indicating possible solutions in each interval. This fact is established by minimizing s subject to (22) within each interval. In both cases, $s_{\min} < 0$, suggesting that $\hat{f}^{\pm}(x)$ must move towards each other to reduce the feasibility region to a point (see Figures 7 and 8). Both intervals must be explored further.

So we subdivide again, and look at the intervals $[0,1]$, $[1,2]$, $[2,3]$ and $[3,4]$. The underestimators $\hat{f}^{\pm}_{[n,n+1]}(x)$ are plotted in Figure 9. For the intervals $[0,1]$, $[1,2]$, and $[3,4]$, the story is the same: $s = 0$ yields feasible points, $s_{\min}$ is negative, and so we must subdivide those intervals further. But something new happens for $[2,3]$. The convex envelope $\hat{f}^{\pm}_{[2,3]}(x)$ completely isolates $f(x)$ from the x-axis. The lower bounding problem (22) is *infeasible* for $s = 0$. The region between $\hat{f}^{+}_{[2,3]}(x)$ and $-\hat{f}^{-}_{[2,3]}(x)$ must be *expanded* before it touches the x-axis (see Figure 10), and thus $s_{\min}$ will be greater than zero. We have rigorously concluded that no solution to $f(x) = 0$ can exist in the interval $[2,3]$, and so we do not need to explore this interval any further. The ability to fathom regions like this is what distinguishes αBB from a straight grid-search.

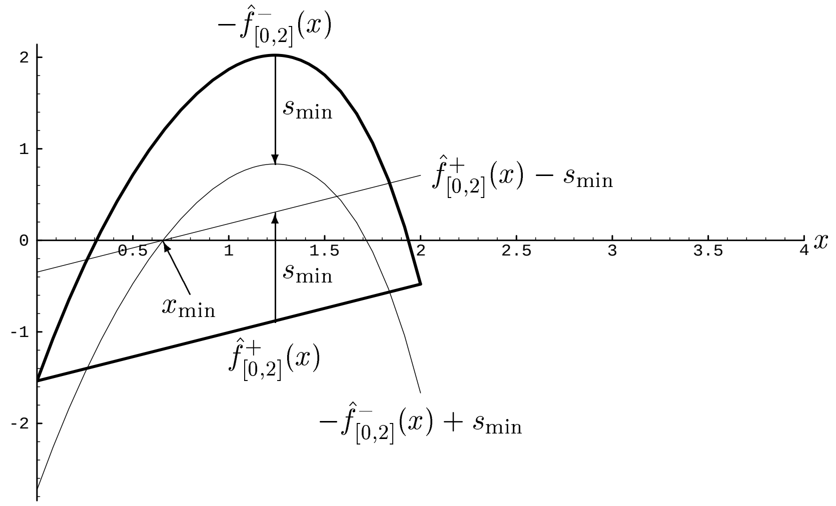

Figure 7: This figure represents the solution to the lower bounding problem in the interval $[0, 2]$. $(x_{\min}, s_{\min}) = (0.656, -1.189)$.

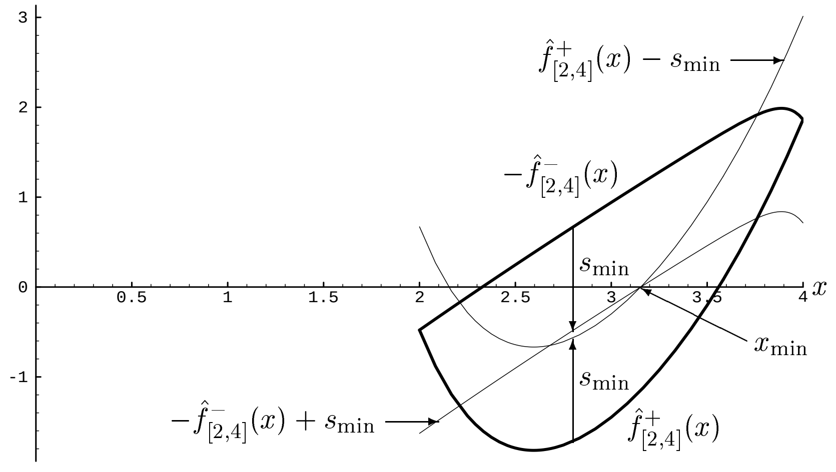

Figure 8: This figure represents the solution to the lower bounding problem in the interval $[2, 4]$. $(x_{\min}, s_{\min}) = (3.154, -1.150)$.

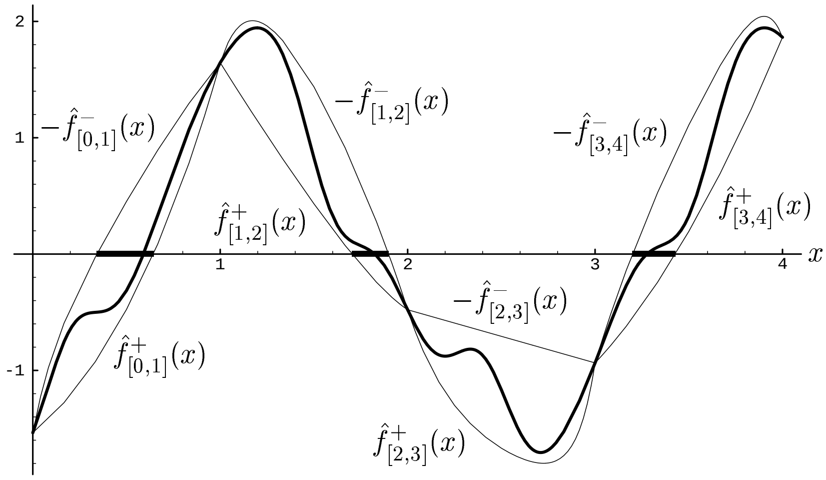

Figure 9: The intervals $[0, 2]$ and $[2, 4]$ have been further subdivided into $[0, 1]$, $[1, 2]$, $[2, 3]$, and $[3, 4]$. Shown above are the convex envelopes around $f(x)$ formed by convex underestimators in each of these intervals. Note that the convex envelopes for $[0, 1]$, $[1, 2]$, and $[3, 4]$ intersect the x-axis, but the convex envelope for $[2, 3]$ does not. This will allow us to conclude rigorously that no solutions to $f(x) = 0$ exist in $[2, 3]$.

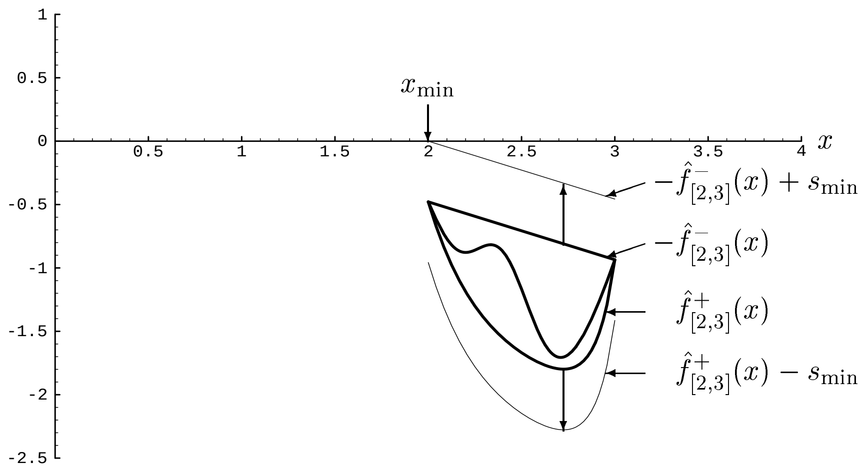

Figure 10: The lower bounding problem for the interval $[2, 3]$ is solved. Note that the convex envelope must be *expanded* before it touches the x-axis resulting in a positive value for $s_{\min}$. This interval will be fathomed. $(x_{\min}, s_{\min}) = (2, +0.479)$.

Exploration will continue with the intervals $[0, 1]$, $[1, 2]$, and $[3, 4]$. These intervals will be subdivided and further tested. As the algorithm progresses, most intervals will eventually be fathomed. A few intervals (three, in fact) will survive. Each of these intervals surrounds a solution point, which will be located by a local search once the interval size is small enough.

III Structure Prediction of Polypeptides

A Structure Prediction of Oligopeptides

The use of computational techniques and simulations in addressing the protein folding problem became possible through the introduction of qualitative and quantitative methods for modeling these systems. Given a sufficiently accurate description of the intramolecular forces, it is in principle possible to predict the folded conformation by optimization. In our work, we have focused not only on the development of global optimization methods, but also on the verification of energy modeling techniques.

In the area of energy modeling, our work has involved the investigation of numerous detailed representations of protein systems. In addition to the traditional all-atom potential energy models, our work has explored the effects of solvation contributions. In fact, although the problem of considering solvation effects in global conformational energy searches has been made tractable by the development of implicit solvation models, results for such formulations are essentially nonexistent, and those that have appeared are for limited searches only. In our work, both solvent accessible area and volume effects have been considered in the context of global searches for oligopeptides. In addition, we have examined the effects of several parameterizations for these models, and have been able to identify those that provide the best correspondence between computational and experimental results.

1 Potential energy models

There are a number of approaches that may be used to model protein interaction energies. In reality, the dynamics of atoms are governed by the quantum theory of their participating electrons. Using the Born-Oppenheimer approximation, one can determine the energy for fixed atomic nuclei from the smallest eigenvalue of the Hamiltonian of the electron system. These approximations and their derivatives are calculated using ab–initio methods. However, due to their computational complexity, such calculations are limited to extremely small molecules. Less detailed, semi-empirical methods are based on all atom representations of the peptide. In general, these models, also known as force fields, are expressed as summations of empirically derived potential functions, with the mathematical form of individual energy terms based on the phenomenological nature of that term. Other simplified models have been used to reduce the degrees of freedom associated with the conformational energy expressions.

A number of empirically-based molecular mechanics models have been developed for protein systems, including AMBER [113, 243, 244], CHARMM [31], DISCOVER [53], ECEPP [171, 172, 173], ECEPP/2 [181], ECEPP/3 [180], ENCAD [51, 149], GROMOS [232], MM2 [7], and MM3

[8, 151, 152]. A general total energy equation, such as Equation 23, includes terms for bond stretching (E_{bond}), angle bending (E_{angle}), torsion (E_{tor}), nonbonded (E_{nb}) and coupled (E_{cross}) interactions.

$$E_{tot} = E_{bond} + E_{angle} + E_{tor} + E_{nb} + E_{cross} \tag{23}$$

Bond stretching and angle bending energies are included in those force fields that allow flexible geometries. A simple representation for both terms is based on the harmonic approximation, which corresponds to the classical description of the movement of a spring (by Hooke's law). The simplest approach, based on the fact that most bonds are near the minimum of their respective energy well, employs a quadratic term to model bond stretching and angle bending energies, as shown in Equations 24 and 25.

$$E_{bond} = \frac{k_{bond}}{2}(l - l_o)^2 \tag{24}$$

$$E_{angle} = \frac{k_{angle}}{2}(\theta - \theta_o)^2 \tag{25}$$

These equations act as penalty functions to force bond distances and bond angles, l and θ, to reference bond lengths and distances, l_o and θ_o, whose values depend on the specific atoms involved. In actuality, these energy terms are more complicated. For bond energies cubic terms are often introduced, and angle energy terms usually include higher power expansions.

Torsional terms are used to describe the internal rotation energy of torsion angles, which exist between all atoms with a 1–4 relationship (separated by two other atoms). For rigid geometry force fields, these torsion angles can be used to define a set of independent variables that effectively describe any protein conformation. This approximately reduces the number of variables by a factor of 10 over those force fields that use a Cartesian coordinate system to describe flexible molecular geometries. In addition, bond and angle energies can be neglected for rigid geometry force fields. The torsion energy expression is typically represented by a Fourier series expansion which, as shown in Equation 26, includes three terms.

$$E_{tor} = E_1(1 - cos\phi) + E_2(1 - cos2\phi) + E_3(1 - cos3\phi) \tag{26}$$

The parameters involved in this expansion, namely E_1, E_2 and E_3, are torsional barriers that are usually specified for the pair of atoms around which the torsion occurs. Each term can be interpreted physically. The 1-x ($cos\phi$) symmetry term accounts for those nonbonded interactions not included in general nonbonded terms. The 2-x ($cos2\phi$) symmetry term is related to the interactions of orbitals, while the 3-x ($cos3\phi$) symmetry term describes steric contributions.

Nonbonded energy terms attempt to model electrostatic and van der Waals interactions between those atoms that are not connected to each other or through a common atom. Typically, a Coulombic term is used to represent electrostatic energies based on atomic point charges, as shown in Equation 27.

$$E_{elec} = \frac{Q_i Q_j}{\epsilon R_{ij}} \tag{27}$$

Here Q_i and Q_j represent the two point charges, while R_{ij} equals the distances between these two points. In some force fields, Coulombic interactions are modified by changing the dependence of the dielectric constant, ϵ. In general, van der Waals interactions are modeled using a 6-12 Lennard-jones potential energy term. This expression, shown in Equation 28, consists of a repulsion and attraction term.

$$E_{vdw} = \epsilon_{ij}\left[\left(\frac{R_{ij}^*}{R_{ij}}\right)^{12} - 2\left(\frac{R_{ij}^*}{R_{ij}}\right)^{6}\right] \tag{28}$$

The energy minimum for a given atomic pair is described by the potential depth, ϵ_{ij}, and position, R_{ij}^*. Other force fields model van der Waals interactions using a modified Hill equation, which replaces the twelfth power term in Equation 28 with an exponential term [7, 8]. Different approaches are also used to describe nonbonded interactions between those atoms that may form hydrogen bonds. Some force fields model these interactions using only Coulombic terms, while other force fields employ special functions, such as a modified 10-12 Lennard-jones type potential term [162], as shown in Equation 29.

$$E_{hbond} = \epsilon_{ij}\left[5\left(\frac{R_{ij}^*}{R_{ij}}\right)^{12} - 6\left(\frac{R_{ij}^*}{R_{ij}}\right)^{10}\right] \tag{29}$$

The cross term, shown in Equation 23, accounts for interactions due to the inherent coupling between bonds, angles and torsions. Generally, these terms are small, and in many force fields they are neglected. Correction terms, which vary for each force field, are also typically added to the general energy equation. For example, the formation of disulfide bridges can be enforced by adding a penalty term to constrain the values of specified bond angles and bond lengths. Correction terms have also been used to adjust conformational energies according to the configurations of proline and hydroxyproline residues [180].

For a significant portion of this work, the ECEPP/3 (Empirical Conformational Energy Program for Peptides) [180] potential model is utilized. In this force field, it is assumed that the covalent bond lengths and bond angles are fixed at their equilibrium values. Then, the conformation is only a function of the independent torsional angles of the system, also known as dihedral angles. The total conformational energy is calculated as the sum of the electrostatic, nonbonded, hydrogen bonded, and torsional contributions. There is also a pseudo–potential for loop closing if the polypeptide contains two or more sulfur–containing residues. More recent work includes a revised treatment of prolyl and hydroxyprolyl residues [180]. For each prolyl or hydroxyprolyl residue contained in the polypeptide a fixed internal conformational energy for the pyrolidine ring is added. The main energy contributions (electrostatic, nonbonded, hydrogen bonded) are computed as the sum of terms for each atom pair (i,j) whose interatomic distance is a function of at least one dihedral angle. The general potential energy terms of ECEPP/3 are shown in Figure 11, while the development of the appropriate parameters is discussed and reported elsewhere [180].

$$
\begin{aligned}
E = & \sum_{(i,j)\in ES} 332.0\frac{q_i\, q_j}{D r_{ij}} && \text{(Electrostatic)}\\
& + \sum_{(i,j)\in NB} F\left(\frac{A}{r_{ij}^{12}} - \frac{C}{r_{ij}^{6}}\right) && \text{(Nonbonded)}\\
& + \sum_{(hx)\in HX} F\left(\frac{A^{\prime}}{r_{hx}^{12}} - \frac{B}{r_{hx}^{10}}\right) && \text{(Hydrogen bonded)}\\
& + \sum_{k\in TOR} \left(\frac{E_0}{2}\right)(1 \pm \cos n_k\theta_k) && \text{(Torsional)}\\
& + \sum_{l\in LOOP} B_L \sum_{il=1}^{il=3} (r_{il} - r_{io})^2 && \text{(Cystine Loop-Closing)}\\
& + \sum_{l\in LOOP} A_L (r_{4l} - r_{4o})^2 && \text{(Cystine Torsional)}
\end{aligned}
$$

Figure 11: 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}, $A^{\prime}_{ij}$ and B_{ij} are nonbonded and hydrogen bonded parameters specific to the atomic pair. $E_{o,k}$ and $E_{o,l}$ 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. The cystine loop–closing term is calculated as a penalty term of three distances involved in loop–closing, where r_{il} represents the actual distance and r_{io} represents the required distance. B_i, the penalty parameter, is set equal to 100. Finally, E_p is a fixed internal energy that is added for each proline residue in the protein. Energy units are kcal/mole and distance units are $\mathring{A}$

2 Solvation energy models

Solvation contributions are generally believed to be a significant force in stabilizing the native conformations of proteins. Explicit methods can be used to include solvation effects by actually surrounding the polypeptide with solvent molecules and calculating solvent–peptide and solvent–solvent interactions. Although these methods are conceptually simple, explicit inclusion of solvent molecules greatly increases the computational time needed to simulate the polypeptide system. Therefore, most simulations of this type are limited to restricted conformational searches. In addition, it is difficult to quantify the effect of hydrophobic interactions that result from the ordering of water molecules.

Methods for estimating solvent free energies have also been developed using both integral equations and continuum models. Integral equation methods can be used to evaluate solvent structure and thermodynamic properties. Typically, molecular dynamics or Monte Carlo simulations are used to calculate ensemble averages from which free energy differences can be obtained. A number of methods have been proposed to estimate these solvation free energies from simulations based on molecular dynamics and Monte Carlo averages [55, 134, 223]. The integral equation method has also been used to analyze the solvent structure of a protein system [127]. In contrast, continuum models use a simplified representation of the solvent environment by neglecting the molecular nature of the water molecules. Calculations of solvation free energies using electrostatic continuum models rely on numerical solutions to the Poisson–Boltzmann equation from which dielectric and ionic strength effects are obtained [101]. Other continuum models estimate free energies of solvation as a function of surface areas and volumes.

In this work, solvation contributions are included implicitly using empirical correlations with both surface area [191] and volume [14]. The main assumption of these models is that, for each functional group of the peptide, a hydration free energy can be calculated from an averaged free energy of interaction of the group with a layer of solvent known as the hydration shell. In addition, the total free energy of hydration is expressed as a sum of the free energies of hydration for each of the functional groups of the peptide; that is, an additive relationship is assumed.

Accessible surface area methods assume that the free energy of hydration is proportional to the solvent–accessible surface area of the peptide, as described by the following equation:

$$E_{HYD} = \sum_{i=1}^{N} (A_i)(\sigma_i) \tag{30}$$

In Equation (30), an additive relationship for N individual functional groups is assumed. (A_i) represents the solvent–accessible surface area for the functional group, and (σ_i) is an empirically derived free energy density parameter.

There are a number of ways to define the surface of a peptide. In developing these surfaces the peptide is represented by a union of spheres, with the radii of the spheres set by the van der Waals radii of the constituent atoms. A spherical test probe is then rolled over these spheres, thereby tracing out a surface. The molecular surface is set by direct contact between the probe sphere and the peptide spheres. In areas where the probe cannot make direct contact, the closest part

Table 1: Free energy density of solvation parameters for the ASP set employed with the solvent accessible surface area model. The first column describes the atom type, whereas the remaining columns provide the solvation parameters in cal/(mol $\mathring{A}^2$) for the corresponding ASP set.

Atom Type	**WE1**	**WE2**	**OONS**	**SCKS**	**JRF**
C aliphatic	12.0	4.0	8.0	32.5	216.0
C carboxyl, carbonyl	12.0	4.0	427.0	32.5	-732.0
C aromatic	12.0	4.0	-8.0	32.5	-678.0
N noncharged	-116.0	-113.0	-132.0	-17.5	-312.0
N charged	-186.0	-169.0	-132.0	-217.5	-312.0
O carboxyl, carbonyl	-116.0	-113.0	-38.0	-17.5	-262.0
O hydroxyl	-116.0	-113.0	-172.0	-17.5	-910.0
O charged	-175.0	-169.0	-38.0	-280.0	-910.0
S all	-18.0	-17.0	-21.0	-9.0	-281.0

of the probe is used. The solvent–accessible surface is defined by the surface traced by the center of the probe as the probe rolls over the peptide spheres. These areas depend on the radius of the probe sphere; when this radius is set to zero both the molecular and solvent–accessible surface areas become the van der Waals surface of the peptide.

Solvent–accessible surface areas are calculated using the MSEED [191] program, which employs algorithms developed by Connolly [46]. MSEED eliminates many unnecessary computations by considering only those convex faces that are on the accessible surface. Rigorous implementation of Connolly's method requires the calculation of interior surface areas, which are ultimately found to be zero. A full description of the MSEED algorithm is given elsewhere [191]. A number of other methods for calculating surface areas are also available [237, 67, 66].

One potential problem that may arise when calculating accessible surface areas is the appearance of gradient discontinuities. This may occur when a new vertex or edge appears on the surface. If the discontinuity is large, minimization techniques requiring gradients may fail to converge to the local minimum conformation. A complete analysis of all situations for which the gradient of the molecular surface area becomes discontinuous has been reported [241].

Once the solvent–accessible surface areas have been calculated, these values must be multiplied by the appropriate (σ_i) parameters as shown in Equation (30). A variety of parameter sets have been developed to model the transfer of atoms from a gaseous to a hydrated environment. The parameter values for the five ASP sets used in this study are given in Table 1.

The ASP sets WE1 and WE2, are taken from Table 3 of [245]. These parameters are both derived from Wolfenden's measured free energies of transfer of amino acid side-chain analogs from vapor to water [251]. Both sets have been adjusted to correct for entropy of mixing effects based on solute and solvent size differences [138, 117], although the applicability of these corrections has been criticized [116, 22]. The parameters for these two sets are negative for all atoms excluding carbon. Qualitatively, this means that the nitrogen, oxygen and sulfur atoms are considered hydrophilic, that is they favor solvent exposure. Comparatively, the WE1 and WE2 parameters are similar, with the largest relative difference being a 3:1 ratio (WE1:WE2) for the σ_C parameter. Therefore, the hydrophobic character of these carbon atoms is stronger for the WE1 ASP set.

The OONS parameter set has been specifically developed to supplement the ECEPP/2 force field [185]. These parameters were derived by a least squares fitting to experimental free energies of gas to water transfer of small aliphatic and aromatic molecules. The most significant difference from the two previous ASP sets is a substantial increase in hydrophobic character for carboxyl (carbonyl) carbon atoms, which corresponds to a hundred-fold increase when compared to the same WE2 parameter. In addition, the free energy parameter becomes negative for aromatic carbons, which indicates a hydrophilic tendency. The three-fold decrease of the OONS values for carboxyl (carbonyl) and charged oxygen atom parameters, as compared to both the WE1 and WE2 ASP sets, is also significant.

Unlike the aforementioned models, the SCKS ASP set is not directly based on experimental free energies [207]. Instead, it is an optimized parameter set developed to complement the CHARMM [31] molecular mechanics force field. Specifically, through the use of experimental and molecular dynamics information, the relative weightings of solvation parameters were refined to provide the best correspondence between minimized and experimental structures. In comparing the individual free energy parameters, it is evident that the hydrophobic character of the carbon atoms is increased approximately three- and eight-fold over the WE1 and WE2 values, respectively. In contrast, the uncharged oxygen and nitrogen atom parameters are 6.5 times smaller (less hydrophilic) than those for the WE1 and WE2 ASP sets. This decrease does not apply to charged oxygen and nitrogen atoms, which possess extremely hydrophilic values.

The JRF ASP set was derived from NMR studies of low energy solvated configurations of 13 tetrapeptides [248]. This represents an important difference from other derivations because actual peptides, rather than simple model compounds, were used to develop the JRF parameters. An ensemble of low energy structures for these tetrapeptides was also produced using the ECEPP/2 potential function. Then, a nonlinear least–squares system was optimized for the best set of atomic solvation parameters. Although, the parameters for oxygen, nitrogen, and sulfur atoms are negative, their large absolute values indicate much larger hydrophilicities than corresponding atoms of any other ASP set. In addition, both the carboxyl (carbonyl) and aromatic carbon atoms possess strong hydrophilic parameters, which contradicts other free energy parameter values for these atoms. The single positive value belongs to the aliphatic carbon atom type, which, although larger than any other parameter for this atom type, possesses the smallest magnitude for the JRF ASP set. Furthermore, because it was developed from minimum energy conformations of peptides, the JRF ASP set has been shown to produce undesirable perturbations during local minimizations if the solvation energy contributions are added at every iteration. Therefore, unlike the aforementioned ASP sets, the JRF solvation energy effects are only included at local minimum conformations.

For volume shell models, the free energy of hydration is assumed to be proportional to the water–accessible volume of a hydration layer surrounding the peptide. This can be represented in the form:

$$E_{HYD} = \sum_{i=1}^{N} (VHS_i)(\delta_i) \tag{31}$$

An additive relationship for the N individual atoms of the peptide is assumed, and (VHS_i) represents the solvent–accessible volume of hydration shell for each atom i which is exposed to water. The (δ_i) parameters are empirically determined free energy of hydration densities for these atoms.

The hydration shell is defined by the volume inside a sphere of radius R_i^h but outside a sphere of radius R_i^v, with both radii centered on atom i. The larger radius, R_i^h, corresponds to the radius of the first hydration shell of atom i, while R_i^v is equal to the van der Waals radius. In order to calculate (VHS_i), the volume of a collection of overlapping hard spheres must be computed using:

$$V(\mathbf{R}) = \sum_i a_i S_i - \sum_{ij} b_{ij} D_{ij} + \sum_{ijk} c_{ijk} T_{ijk} - \sum_{ijkl} d_{ijkl} Q_{ijkl} \tag{32}$$

In Equation (32), S_i signifies the volume of a single sphere, while D_{ij}, T_{ijk} and Q_{ijkl} represent the volume of intersection of two, three and four spheres, respectively. This is sufficient because all higher order overlaps can be decomposed into the three types of intersections included in Equation (32). Therefore, the solvent–accessible volume of hydration can be written as:

$$(VHS_i) = V(R_i^h) - V(R_i^v) \tag{33}$$

The first term in Equation (33) is calculated using Equation (32) with the radii of all atoms set equal to their van der Waals radii, while the second term is calculated with the radius of atom i equal to R_i^h and the van der Waals radii of all the other atoms. A number of methods to compute hydration shell volumes have been proposed [14, 102, 121].

The form of Equation (32) is not suitable for force field models using pairwise intramolecular potential, such as ECEPP/3. Furthermore, direct truncation at the double–overlap term would lead to large errors. In this work, the RRIGS (Reduced Radius Independent Gaussian Sphere) approximation is used to efficiently calculate the exposed volume of the hydration shell [14]. This method uses a truncated form of Equation (32) but also artificially reduces the van der Waals radii of all atoms other than atom i when calculating (VHS_i). These reductions effectively decrease the contribution of the double overlap terms, leading to a cancellation of the error which results from neglecting the triple and higher overlap terms. In addition, the characteristic density of being inside the overlap volume of two intersecting spheres is not represented as a step function, but as a Gaussian function; this provides continuous derivatives of the hydration potential. Therefore, the solvation energy contributions can easily be added at every step of local minimizations since the RRIGS approximation has the same set of interactions as the ECEPP/3 potential.

Free energy density parameters for solvent accessible volumes have been developed for nonionic and charged organic solute molecules [122, 123, 120]. In this work, RRIGS specific (δ_i) parameters, which were developed by a least square fitting of experimental free energy of solvation data for 140 small organic molecules [14], are used (Table 2). The classification of the RRIGS atom types is more fragmented than for the solvent accessible surface area ASP sets. The most hydrophilic values belong to the nitrogen and selected oxygen and hydrogen atom types. In addition, aromatic carbons tend to possess slightly hydrophilic values, whereas the carbonyl and aliphatic carbon atoms exhibit the most hydrophobic parameter values.

Table 2: Free energy density of solvation parameters employed in the RRIGS model. The second column provides the solvation parameters in cal/(mol $\AA^2$), and the last two columns correspond to the van der Waals and hydration radii ($\AA$), respectively.

Atom Type	δ	R^v	R^h
H hydroxyl, amino	-10.35	1.415	4.17
H acid	-3.206	1.415	4.17
H amide	-7.714	1.415	4.17
H thiol	2.709	1.415	4.17
C aliphatic CH_3	1.319	2.125	5.35
C aliphatic CH_2	0.2374	2.225	5.35
C aliphatic CH	-1.271	2.375	5.35
C other aliphatic	-2.297	2.060	5.35
C cyclic CH	0.2890	2.375	5.35
C aromatic CH	-0.2137	2.100	5.35
C aromatic CR	-1.713	1.850	5.35
C branched aromatic C	-1.910	1.850	5.35
C aromatic COH	-0.6063	1.850	5.35
C carbonyl	2.696	1.870	5.35
N primary amine	-1.149	1.755	5.05
N secondary amine	-10.28	1.755	5.05
N aromatic	-10.48	1.755	5.05
N amide	-7.332	1.755	5.05
O hydroxyl, ether	-7.396	1.620	4.95
O acid, ester	0.07897	1.620	4.95
O ketone, carbonyl	-15.70	1.560	4.95
O acid, amide carbonyl	-15.56	1.560	4.95
S thiol, disulfide	-4.706	2.075	5.37

3 Global optimization framework

The energy minimization problem is formulated as a unconstrained nonconvex global optimization problem, which is fashioned after the general formulation given in problem 1. Let $i = 1, \ldots, N_{RES}$ be an indexed set describing the sequence of amino acid residues in the peptide chain. There are $\phi_i, \psi_i, \omega_i,\ i = 1, \ldots, N_{RES}$ dihedral angles along the backbone of this peptide. In addition, let K^i denote the number of dihedral angles for the side chain of the i^{th} residue and J^N and J^C denote the number of dihedral angles for the amino and carboxyl end groups, respectively. Using these definitions the optimization problem takes the following form:

$$\min \qquad E(\phi_i, \psi_i, \omega_i, \chi_i^k, \theta_j^N, \theta_j^C) \tag{34}$$

$$\begin{array}{rrcll} \text{subject to} & -\pi & \le & \phi_i \le \pi, & i = 1, \ldots, N_{RES} \\ & -\pi & \le & \psi_i \le \pi, & i = 1, \ldots, N_{RES} \\ & -\pi & \le & \omega_i \le \pi, & i = 1, \ldots, N_{RES} \\ & -\pi & \le & \chi_i^k \le \pi, & i = 1, \ldots, N_{RES},\ k = 1, \ldots, K^i \\ & -\pi & \le & \theta_j^N \le \pi, & j = 1, \ldots, J_N \\ & -\pi & \le & \theta_j^C \le \pi, & j = 1, \ldots, J_C \end{array}$$

In general, E represents the total potential energy function and the free energy of solvation. However, in the case of the JRF ASP set, the potential energy function is minimized before adding the hydration energy contributions for this ASP set. In other words, gradient contributions from solvation are not considered. This approach is represented by the following equation:

$$E_{JRF}^{Total} = E_{Min}^{Unsol} + E_{JRF}^{Sol} \tag{35}$$

Even after reducing this optimization problem to a function of internal variables (dihedral angles), the multidimensional surface that describes the energy function has an astronomically large number of local minima. A large number of techniques have been developed to search this nonconvex conformational space. In general, the major limitation is that these methods depend heavily on the supplied initial conformation. As a result, there is no guarantee for global convergence because large sections of the domain space may be bypassed. To overcome these difficulties, the αBB global optimization approach [1, 4, 2, 3, 9] has been extended to identifying global minimum energy conformations of solvated peptides. The αBB global optimization algorithm effectively brackets the global minimum solution 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 energy function, E. Lower bounds belong to the set of solutions of the convex lower bounding functions, which are constructed by augmenting E with the addition of separable quadratic terms. The lower bounding functions, L, of the energy hypersurface can be expressed in the following manner:

$$
\begin{aligned}
L \;=\; E + \quad & \sum_{i=1}^{N_{RES}} \alpha_{\phi,i}\left(\phi_i^L - \phi_i\right)\left(\phi_i^U - \phi_i\right) + \\
& \sum_{i=1}^{N_{RES}} \alpha_{\psi,i}\left(\psi_i^L - \psi_i\right)\left(\psi_i^U - \psi_i\right) + \\
& \sum_{i=1}^{N_{RES}} \alpha_{\omega,i}\left(\omega_i^L - \omega_i\right)\left(\omega_i^U - \omega_i\right) + \\
& \sum_{i=1}^{N_{RES}} \sum_{k=1}^{K^i} \alpha_{\chi,i,k}\left(\chi_i^{k,L} - \chi_i^k\right)\left(\chi_i^{k,U} - \chi_i^k\right) + \\
& \sum_{j=1}^{J^N} \alpha_{j,\theta^N}\left(\theta_j^{N,L} - \theta_j^N\right)\left(\theta_j^{N,U} - \theta_j^N\right) + \\
& \sum_{j=1}^{J^C} \alpha_{j,\theta^C}\left(\theta_j^{C,L} - \theta_j^C\right)\left(\theta_j^{C,U} - \theta_j^C\right)
\end{aligned} \tag{36}
$$

Here $\phi_i^L, \psi_i^L, \omega_i^L, \chi_i^{k,L}, \theta_j^{N,L}, \theta_j^{C,L}$ and $\phi_i^U, \psi_i^U, \omega_i^U, \chi_i^{k,U}, \theta_j^{N,U}, \theta_j^{C,U}$ represent lower and upper bounds on the dihedral angles $\phi_i, \psi_i, \omega_i, \chi_i^k, \theta_j^N, \theta_j^C$. 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 over the defined domain. The computational requirement of the αBB algorithm depends on the number of variables (global) on which branching occurs. Therefore, these global variables need to be chosen carefully.

The determination of the global minimum energy conformation using αBB requires the interfacing of a number of programs (αBB [1, 4, 2, 3, 9], PACK [205], NPSOL [79] and potential and solvation energy modules). PACK, a peptide generation program, is called once directly by αBB in order to initialize the current problem. In subsequent steps PACK is called through NPSOL [79], a local nonlinear optimization solver used to solve both the upper and lower bounding problems. PACK internally transforms to and from Cartesian and internal coordinate systems, and provides potential energy and gradient contributions for the ECEPP/3 potential model at every step of the local minimizations. When considering surface–accessible solvation, surface areas are calculated using MSEED [191]; whereas volumes of hydration shells are determined using the RRIGS module [14]. Finally, an additional module, UBC (Upper Bound Check), is used to verify the quality of the upper bound solutions. The entire suite of programs has been combined to form the GLO-FOLD software package for the prediction of protein structure, as shown in Figure 12.

The basic steps of the algorithm are as follows:

(1) The initial best upper bound is set to an arbitrarily large value. The original domain is partitioned along one of the global variables.

(2) A convex function (L) is constructed in each hyper–rectangle and minimized using NPSOL, with calls (through PACK) to both ECEPP/3 and one of the two solvation modules. If a

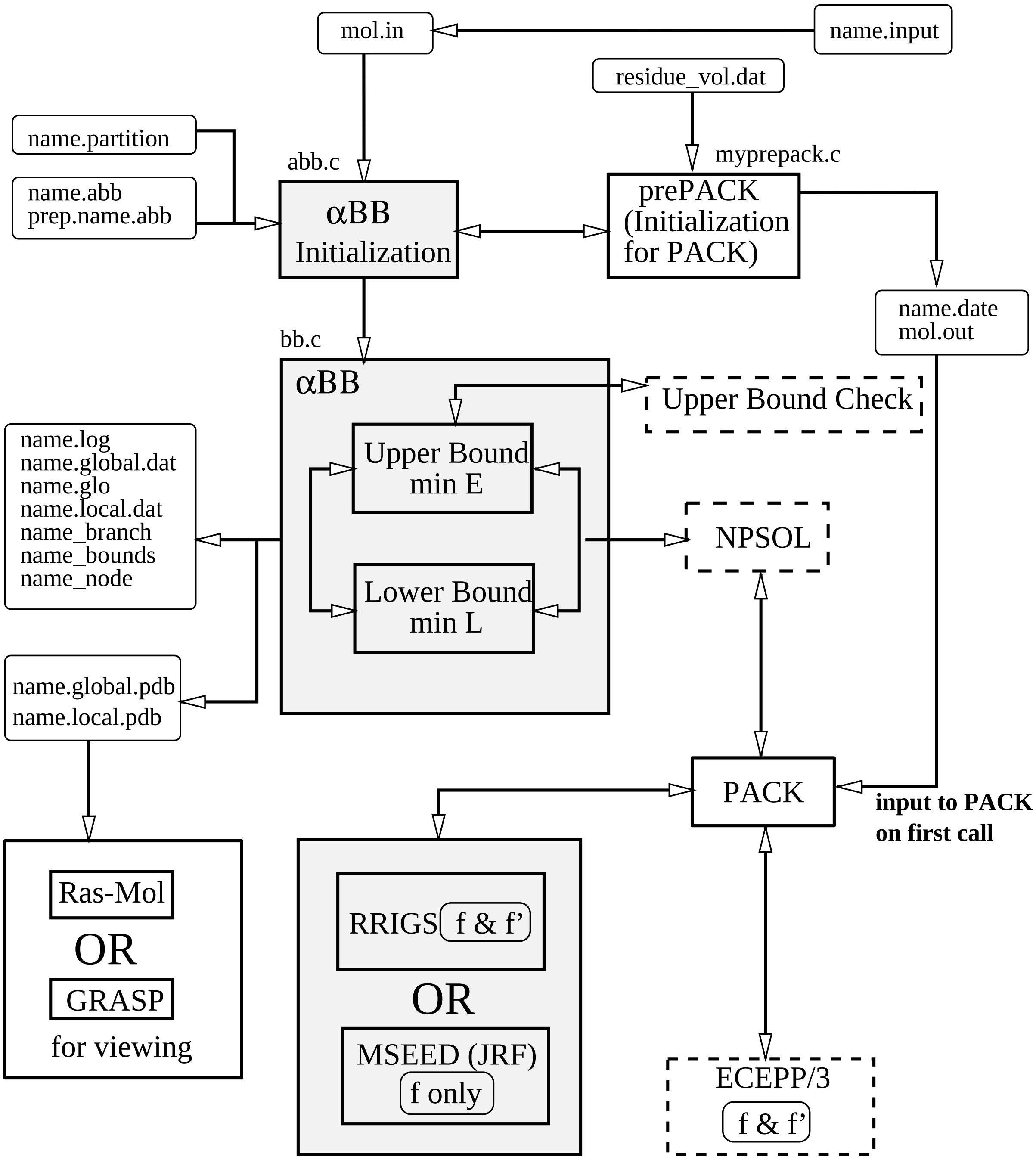

Figure 12: Interface for αBB within GLO-FOLD. The arrows indicate the direction of information flow. The names of the input, output, and intermediate files are indicated, in addition to selected source code files. References to "f & f'" and "f only" describe whether gradient evaluations or only function evaluations are used in the respective modules.

solution is greater than the current best upper bound the entire subregion can be fathomed, otherwise the solution is stored.

(3) The local minima solutions for L are used as initial starting points for local minimizations of the upper bounding function (E) in each hyper–rectangle. Again, the appropriate calls are made to PACK and the potential and solvation energy modules. In solving the upper bounding problems, all variable bounds are expanded to [-180,180]. These solutions are upper bounds on the global minimum solution in each hyper–rectangle.

(4) The current best upper bound is updated to be the minimum of those thus far stored. If a new upper bound (from step 3) is selected, the upper bound check, UBC, module is called. UBC checks that the absolute value of each gradient in the objective function gradient vector is below a specified tolerance (kcal/mole/deg). If a gradient does not satisfy this check the corresponding variable bounds are incrementally increased and the problem is solved with the previous point used as the initial starting point. This process is repeated until the gradient constraints are satisfied or an iteration limit is exceeded. UBC also employs algorithms to calculate the second derivative matrix [183] which is used to verify that the upper bound solution is a local minimum, that is the Hessian matrix is positive semi–definite. If the matrix is not positive semi–definite or the gradient checks are not satisfied, the upper bound solution is rejected.

(5) The hyper–rectangle with the current minimum value for L is selected and partitioned along one of the global variables.

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

4 Computational studies

Single residue examples were defined as terminally blocked by using acetyl (amino) and methyl (carboxyl) end groups. All dihedral angles were treated as global variables, excluding the three θ angles of the end groups. The relative convergence was set to 10^{-2}. For these examples, all dihedral angles, excluding those of the end groups, were treated as global variables. The remaining variables were treated locally; that is, they were allowed to vary during minimization, but their domain space was not partitioned. When using the RRIGS and JRF models the global variables were assigned initial α values of 3.0. For the other solvation models, the α values were increased to 5.0.

For a number of residues, the JRF global minimum solutions possess ω angles in the range of [-30,30] with the corresponding ϕ and ψ angles near the [-150,80] region. Additional runs were conducted in which the ω angles were constrained to the range of [160,200]. In all cases, with the exception of serine, this constraint led to increases in solvation energies and decreases in potential energy terms while the structures became either β–sheet–like or α–helical. Without exception, the ω angles for the all other global minimum energy solutions were within the [160,200] range.

The remaining analysis in this section refers to these constrained (ω within [160,200]) minima for the JRF ASP set. This is appropriate not only in comparing the JRF results with other solvation results, but it also makes the analysis relevant for the oligopeptide studies because similar ω bounds are typically used.

The results of the solvation models are more clearly evaluated when examining energy differences. For example, ΔE^{POT} ($\Delta E^{POT} = E^{POT}_{ASP}$ - E^{POT}_{RRIGS}) refers to the change in potential energy of an area based global minimum (E^{POT}_{ASP}) and the RRIGS global minimum (E^{POT}_{RRIGS}) solution for a given terminally blocked residue. This difference is positive in almost all cases, which indicates that the potential energy of the RRIGS structure is always lower and provides more stabilization at the corresponding global minimum solution. In most cases, this difference is very small, especially for the OONS and SCKS ASP sets. In fact, for both these ASP sets several residues, most noticeably phenylalanine and tyrosine, have more potential energy stabilization at their corresponding global minima. However, for five peptides, namely phenylalanine, serine, threonine, tyrosine and leucine, the JRF potential energy is more than 10 kcal/mole less stabilizing. This set of residues includes the three residues (serine, threonine and tyrosine) that contain hydroxyl groups among the side chain atoms, as well as the two regular aromatic residues (phenylalanine and tyrosine). It is also interesting to note that these atom types (i.e., hydroxyl oxygen and aromatic carbon) correspond to two of the most hydrophilic type atoms in the JRF ASP set. Finally, the leucine ΔE^{POT} seems to be abnormally high because of the large torsional contribution at the JRF global minimum conformation.

The results for ΔE^{HYD} ($\Delta E^{HYD} = E^{HYD}_{ASP}$ - E^{HYD}_{RRIGS}), which refers to the change in hydration energy between an area based global minimum (E^{HYD}_{ASP}) and the RRIGS global minimum (E^{HYD}_{RRIGS}) solutions, are especially interesting. These differences are positive in most cases, which indicates that the hydration energy of the RRIGS structure is generally lower. However, when examining the JRF results ΔE^{HYD} is negative for four examples, namely histidine, phenylalanine, tryptophan and tyrosine. Excluding the special case of proline, these four residues correspond to the naturally occurring residues which possess ringed side chain structures. Other trends are also apparent. The most positive ΔE^{HYD} values for the JRF ASP set are provided by the aliphatic residues. In addition, the acidic residues, glutamic and aspartic acid, and the amide forms of these residues, glutamine and asparagine, have comparable values for ΔE^{HYD}.

For the other (gradient inclusive) ASP sets the ΔE^{HYD} of different residues are less varied. However, it is important to consider that for all residues, excluding tyrosine, the ASP sets follow a WE2, WE1, OONS and SCKS order when ranked beginning with the most stabilizing hydration energy. Low hydration energies are expected for WE2 because of the consistently small hydrophobic and relatively large hydrophilic parameters. In most cases, the WE1 ΔE^{HYD} are only slightly larger than for WE2. This can be directly attributed to the increased hydrophobicity of the free energy parameter for the carbon atoms of the WE1 ASP set. When comparing the OONS and WE1 ASP sets the increased ΔE^{HYD} is more noticeable, which is most likely a result of the combined effects of the strong hydrophobic value of the carboxyl (carbonyl) carbon parameter and the decreased hydrophilic value of the carboxyl (carbonyl) oxygen parameter for the OONS ASP set. However, for aromatic residues (i.e., phenylalanine, tryptophan and tyrosine) these effects are partially offset by introducing a hydrophilic character for aromatic carbons in the OONS ASP set. In fact, for tyrosine this change is strong enough to cause the OONS ASP set to produce a more stabilizing hydration energy than the WE1 ASP set. A comparison between the OONS and SCKS reveals the largest increase in ΔE^{HYD} values. This can partly be attributed to the relatively large value of the free energy parameters for carbon atoms of the SCKS ASP set. The increase is also enhanced for aromatic residues because of the hydrophilic nature of the aromatic carbon atoms for the OONS ASP set. In addition, for residues with nitrogen containing side chains, the ΔE^{HYD} increase is heightened because of a subsequent decrease in the value of the free energy parameter for nitrogen atoms in the SCKS ASP set. Finally, a comparison of other surface accessible solvation results to the JRF results is qualitatively similar to those made between the RRIGS and JRF models. Specifically, the strong hydration energy stabilization of ring containing residues, as well as the decreased stabilization provided by aliphatic residues, is evident.

A more detailed analysis was performed by generating adiabatically relaxed ϕ–ψ maps for N–acetyl–N'–methyl–alanineamide. The adiabatic curves define regions within a given energy of the global minimum value. The first map corresponds to an adiabatically relaxed map for the unsolvated form of the peptide. This was calculated by fixing the ϕ and ψ angles at 3 degree increments and using a local minimization solver to minimize the ECEPP/3 potential energy by varying the remaining dihedral angles. The other maps were constructed by a similar procedure, although the minimized energy now included both ECEPP/3 and the appropriate hydration free energy. In generating the data for the JRF, the ECEPP/3 energy was first minimized in the absence of solvent at each point and the map was generated by adding the solvation free energy for the JRF model at the minimized conformation.

These maps reveal several important effects of including solvation. Experimental data for the alanine peptide suggests that more than one conformation is present in solution, and NMR coupling constants indicate a large population of conformations with $-70 > \phi > -80$ [156]. It is also expected that hydration may weaken intrapeptide hydrogen bonding. The unsolvated map indicates well defined regions for intramolecular hydrogen bonding (C_7) and for right–handed α–helices (α_R). The global minimum occurs within the C_7 region. The RRIGS map retains some features of the unsolvated map, with the global minimum in the C_7 region and a very strong α_R region. However, there is a broadening of the β–sheet (C_5) region as well as a less distinct C_7 minimum. This can be contrasted with both the WE1 and WE2 adiabatic maps, which exhibit large C_5 regions and significant decreases in the size of both the C_7 and α_R regions. The OONS map contains an even larger low energy region that connects the C_5 and C_7 domains. The α_R low energy region is also broader than either of those indicated by the WE1 or WE2 map. In all three cases (WE1, WE2 and OONS) the global minimum is shifted to the β–sheet domain. In contrast, the SCKS adiabatic map is more similar to the RRIGS map because of its smaller and disjoint C_5 and C_7 regions, as well as the location of the global minimum in the C_7 well. The largest disparity between these maps exists with the JRF adiabatic map, which indicates a complete shift away from the C_7 minimum towards the C_5 region.

Qualitatively, similar trends are observed for the ϕ–ψ distribution of other terminally blocked amino acids. The RRIGS model predicts a majority of global minima in the C_7 region, which indicates a tendency to preserve certain potential energy effects. As expected, the majority of WE1 and WE2 global minima lie within the C_5 domain, with the same distribution for each parameter set. The most uniform distribution of global minima belongs to the OONS ASP set, for which there are an almost equal number of C_5 and C_7 global minimum structures. This agrees with the large low energy regions displayed on the N–acetyl–N'–methyl–alanineamide adiabatic map. The large population of C_7 global minima for the SCKS ASP set is also suggested by the strong C_7 region on the N–acetyl–N'–methyl–alanineamide map. In accordance with the distinct implementation of the JRF model, these results are less predictable. Specifically, although almost half of the JRF global minima lie in the C_5 domain, a significant number also exhibit α-helical type structures, which contrasts with the ϕ-ψ map of N–acetyl–N'–methyl–alanineamide.

Met–enkephalin (H–Tyr–Gly–Gly–Phe–Met–OH) is an endogenous opioid pentapeptide found in the human brain, pituitary, and peripheral tissues and is involved in a variety of physiological processes. The peptide consists of 24 independent torsional angles and a total of 75 atoms, and has played the role of a benchmark molecular conformation problem. The energy hypersurface is extremely complex with the number of local minima estimated on the order of 10^{11} [150]. Based on a previous study, the unsolvated global minimum potential energy conformation, with an ECEPP/3 energy of -11.707 kcal/mole, was shown to exhibit a type II' β–bend along the N–C' peptidic bond of Gly^3 and Phe^4 [10].

In studying the effects of solvation on the structure of met–enkephalin, the results for the unsolvated structure were verified by employing the algorithm outlined in Section 4.2. A major difference from the previous implementation [10] is the addition of the UBC module, as well as the expansion of all variable bounds (to [-180,180]) when solving the upper bounding problems. Since the backbone dihedral angles (i.e., ϕ and ψ) are the most influential variables in defining the backbone structure, the corresponding 10 backbone dihedral angles were treated as global variables for the enkephalin problems. Although they were not partitioned during the global search, all other variables (i.e., ω and all χ) were allowed to vary during local minimizations. The global variables were assigned initial α values of 5.0 when using the unsolvated, RRIGS and JRF models, and values of 10.0 for all other models. In the case of unsolvated met–enkephalin, the structural and energetic results of the previously identified global minimum energy structure [10] were confirmed.

Experimental results have indicated that met–enkephalin in aqueous solution does not possess an unique structure [87]. In general, experimentally determined aqueous conformations are found to exhibit characteristics of extended random–coil polypeptide with no discernible secondary structure. When considering the effects of hydration, the competition for backbone hydrogen bonding (with water), which contributes to the bending of the unsolvated conformation, should result in a more extended structure.

The RRIGS model predicts a more extended structure than the global minimum structure reported for the unsolvated case [10]. In fact, although a slight turn occurs near the N–terminus, the structure possesses no hydrogen bonds (< 2.2 $\mathring{A}$) and an overall end–to–end C^{α} distance of

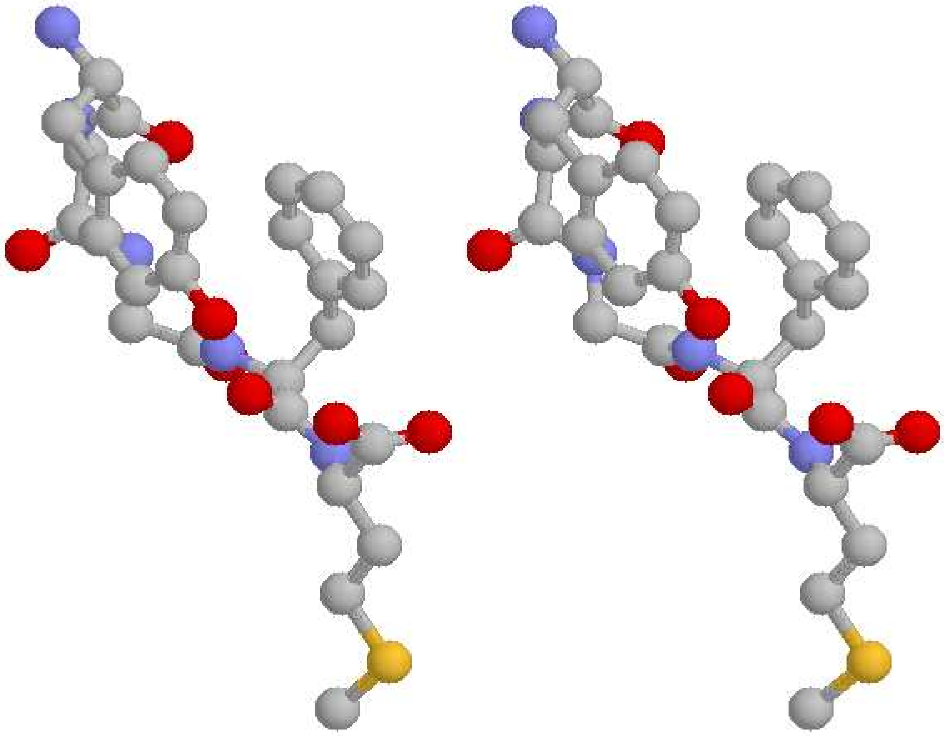

Figure 13: Plot of met–enkephalin conformation (in stereo). Global minimum energy of -50.01 kcal/mole using the RRIGS model for hydration.

10.16 $\AA$. In addition, there exists close proximity of the Tyr and Phe aromatic rings, as shown in Figure 13. The centroids of these rings are separated by 4.16 $\AA$, which is slightly closer than the preferential aromatic–aromatic interaction distance of 4.5 to 7 $\AA$ [33]. Furthermore, the aromatic rings are essentially in a parallel, as opposed to the more common orthogonal, orientation. This suggests an attempt to balance the slightly hydrophilic nature of the aromatic carbon atoms, as given by the RRIGS δ_i, and the favorable hydrophobic interactions between the two rings. The values of the dihedral angles for the global minimum energy conformation are given in Table 3.

The global minimum structures for the area based hydration models (gradient inclusive) are less extended, as exhibited by Figures 14 and 15. The lowest energy structures for the WE1 and WE2 models are very similar, with an the end–to–end C^α distance of 5.85 $\AA$ for both solvation models. In addition, the bend near the N–termini is stabilized by a hydrogen bond between the CO of the tyrosine residue and the NH proton of the phenylalanine residue (approximately 1.98 $\AA$). This bend is similar to the type II' β–bend of the unsolvated global minimum energy structure,

Table 3: Dihedral angles at the global minimum energy conformation of met–enkephalin, using the RRIGS model for hydration.

	ϕ	ψ	ω	χ_1	χ_2	χ_3	χ_4
Tyr	-168.32	-30.81	178.52	-173.58	-101.26	18.83	
Gly	78.83	-86.96	182.73				
Gly	162.94	91.72	172.83				
Phe	-150.72	162.32	181.50	66.66	92.68		
Met	-77.80	106.79	181.63	-67.82	178.91	180.01	-60.01

Table 4: Dihedral angles at the global minimum energy conformation of met–enkephalin, using the WE1 model for hydration.

	ϕ	ψ	ω	χ_1	χ_2	χ_3	χ_4
Tyr	-162.65	-43.34	-177.43	-173.76	-90.62	2.61	
Gly	66.15	-86.62	172.92				
Gly	-152.31	32.40	-178.49				
Phe	-157.59	154.87	179.36	52.02	-96.19		
Met	-90.62	128.89	-179.18	-169.29	176.88	180.14	-59.99

although it is shifted to the Gly^2–Gly^3 backbone region. The aromatic ring separation is wider (approximately 6.48 $\AA$ for both models) than for the RRIGS global minimum structure, although the side chain orientations are similar. The values of the dihedral angles for the WE1 and WE2 global minimum structures are given in Tables 4 and 5 respectively.

The lowest energy conformation for the OONS ASP is also similar to the WE1 and WE2 structures. In this case, the end–to–end C$^\alpha$ distance is again 5.85 $\AA$. The bending near the N–termini is again similar to a type II' β–bend along the Gly^2–Gly^3 backbone, although in this case it is stabilized by a slightly weaker hydrogen bond between the CO of the tyrosine residue and the NH proton of the phenylalanine residue (approximately 2.01 $\AA$). The 6.60 $\AA$ aromatic ring separation is also slightly larger, which may be attributed to the slightly hydrophilic character of

Table 5: Dihedral angles at the global minimum energy conformation of met–enkephalin, using the WE2 model for hydration.

	ϕ	ψ	ω	χ_1	χ_2	χ_3	χ_4
Tyr	-162.70	-43.23	-177.47	-173.94	-90.83	2.63	
Gly	66.15	-86.59	173.03				
Gly	-152.49	32.41	-178.55				
Phe	-157.84	154.97	179.26	52.12	-96.11		
Met	-89.96	129.19	-179.17	-169.47	176.75	180.13	-59.99

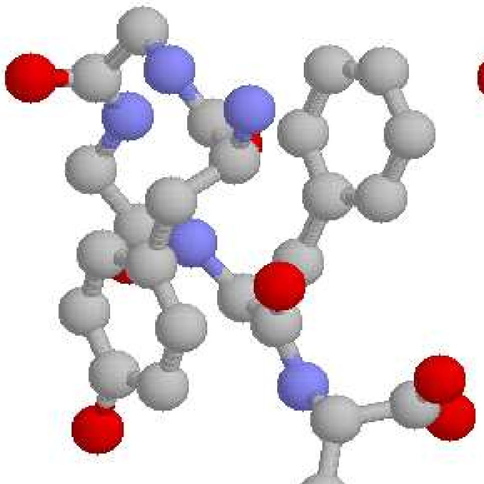

Figure 14: Plot of met–enkephalin conformation (in stereo). Global minimum energy of -30.31 kcal/mole using the WE1 model for hydration.

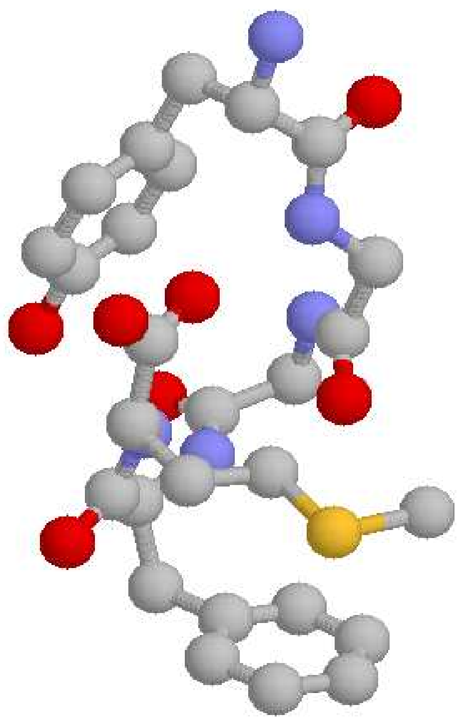

Figure 15: Plot of met–enkephalin conformation (in stereo). Global minimum energy of -0.62 kcal/mole using the SCKS model for hydration.

Table 6: Dihedral angles at the global minimum energy conformation of met–enkephalin, using the OONS model for hydration.

	ϕ	ψ	ω	χ_1	χ_2	χ_3	χ_4
Tyr	-166.11	-50.84	-176.25	-188.97	-102.81	2.45	
Gly	63.86	-86.04	175.39				
Gly	-151.94	33.86	-178.80				
Phe	-159.47	153.41	179.46	50.93	-96.43		
Met	-79.75	148.31	-178.93	-68.16	181.45	178.08	59.70

Table 7: Dihedral angles at the global minimum energy conformation of met–enkephalin, using the SCKS model for hydration.

	ϕ	ψ	ω	χ_1	χ_2	χ_3	χ_4
Tyr	-82.91	154.09	-176.27	-172.88	79.47	-166.08	
Gly	-151.61	81.91	168.71				
Gly	84.09	-72.41	-169.54				
Phe	-137.07	18.52	-173.06	57.94	-86.04		
Met	-162.71	158.63	-179.76	51.94	173.67	179.21	-58.18

the aromatic carbon parameters as compared to the WE1 and WE2 ASP sets. The values of the dihedral angles for the global minimum structure are given in Tables 6.

The SCKS global minimum structure is even less extended, as shown in Figure 15. Although the aromatic ring separation becomes wider (8.13 $\AA$) the overall end–to–end C^{α} distance decreases to 5.80 $\AA$. In this structure, there are two stabilizing hydrogen bonds – a 1.86 $\AA$ hydrogen bond between the NH proton of the first glycine residue and the CO of the methionine residue, and a 2.02 $\AA$ hydrogen bond between the CO of the first glycine residue and the NH proton of the phenylalanine residue. This backbone structure exhibits a type II' β–bend around the Gly^3 and Phe^4 residues, which is similar to the global minimum energy conformation for unsolvated met–enkephalin. This compact structure is consistent with the relatively strong hydrophobic values of all carbon atom free energy parameters, as well as the relatively weak hydrophobic values of the oxygen and nitrogen atoms for the SCKS ASP set. The values of dihedral angles corresponding to the global minimum energy structure are given in Table 7.

In contrast, the JRF global minimum energy structure resembles a more extended conformation, with an overall end–to–end C^{α} distance of 9.56 $\AA$. The plot of this structure, given in Figure 16, shows that the residues near the N–terminus are almost fully extended, although there is slight turn near the C–terminus. This bending is stabilized by the formation of 2.10 $\AA$ hydrogen bond between the CO of the second glycine residue and the NH proton of the methionine residue. In addition, the structure displays a large 14.87 $\AA$ separation between the centroids of the Phe and Tyr aromatic rings. This can be partly attributed to the strongly hydrophilic character of the

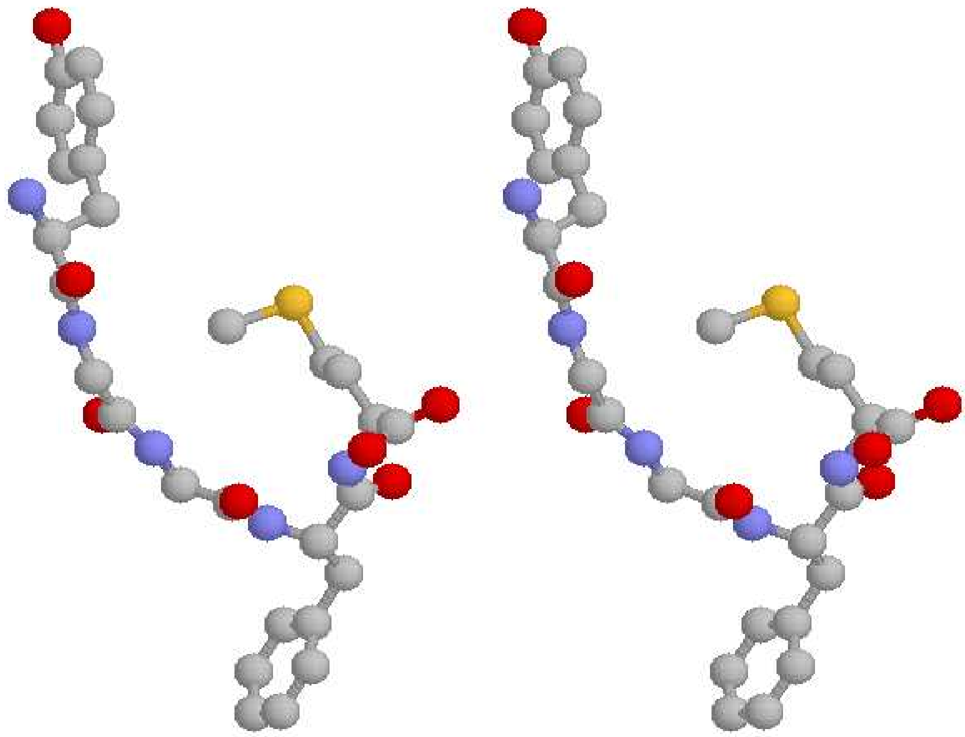

Figure 16: Plot of met–enkephalin conformation (in stereo). Global minimum energy of -283.76 kcal/mole using the JRF model for hydration.

aromatic and carboxyl (carbonyl) carbons parameters for the JRF ASP set. The values of dihedral angles corresponding to the global minimum energy are given in Table 8.

The structures were further analyzed by comparing energy evaluations at corresponding global minimum solutions. This information is given in Tables 9 and 10. In all cases, excluding the SCKS model, the JRF global minimum energy structure provides that most stabilizing values for the hydration energy. However, these stabilizing hydration energies are generally offset by the relatively high value for potential energy at the JRF global minimum conformation (5.06 kcal/mole, obtained by calculating E_{TOT} - E_{HYD} from Tables 9 and 10). In fact, the high potential energy causes the JRF structure to exhibit the highest values for overall energy, excluding the case of the JRF model. Only when considering the JRF model do these stabilizing hydration free energies tend to dominate the prediction of the global minimum structure. This is evidenced by the fact that the JRF structure provides an overall energy, more than 100 kcal/mole lower than any other total energy, which can be directly attributed to the differences in hydration energy. When using the

Table 8: Dihedral angles at the global minimum energy conformation of met–enkephalin, using the JRF model for hydration.

	ϕ	ψ	ω	χ_1	χ_2	χ_3	χ_4
Tyr	-84.96	160.74	179.09	-59.83	100.80	-179.29	
Gly	-160.26	151.83	-177.53				
Gly	159.50	-157.94	178.71				
Phe	-76.55	76.23	-178.05	-61.87	108.68		
Met	-132.90	147.47	-179.83	-65.17	-175.99	-84.91	59.38

SCKS model, the only case for which the JRF conformation does not produce the most stabilizing hydration energy, the JRF structure provides the least stabilizing hydration energy. This indicates that unlike the other hydration models, the SCKS model does not provide more hydration energy stabilization for extended conformations. This agrees with the prediction of the SCKS global minimum energy structure, which exhibits the most folded conformation. The SCKS structure also closely resembles the unsolvated global minimum energy structure and it exhibits the lowest potential energy contribution, -11.63 kcal/mole, which is only .08 kcal/mole higher than the global minimum potential energy. This suggests that low potential energy conformations are not only favored but also enhanced by hydration effects for the SCKS model. Excluding the SCKS model, the other models predict relatively large hydration energies at the SCKS structure. In fact, for the RRIGS, WE1 and WE2 models, the SCKS structure produces the highest values for the hydration energies. For the OONS and JRF model, the hydration energies are only smaller than those for the RRIGS structure. This is consistent with the hydrophilic nature of the aromatic carbons for the OONS and JRF models. Specifically, because the aromatic ring separation is smallest for the RRIGS structure, the OONS and JRF hydration models tend to provide higher hydration energies for this structure. Although hydration energies for the RRIGS structure are typically high, the RRIGS model predicts a stabilizing hydration energy for this structure, second only to the JRF structure. It is this hydration energy contribution, when coupled with a relatively low potential energy (-8.59 kcal/mole), that sets the RRIGS global minimum energy structure. For the other hydration models, low potential energy contributions (-9.77, -9.75 and -9.22 kcal/mole for WE2, WE1 and OONS, respectively) seem to be more important in the prediction of relatively compact structures. In these cases the relative weighting of the hydration energy contributions does not favor extended conformations. However, these models also do not provide low hydration energies at the most compact structures, such as the SCKS global minimum energy structure. This indicates an interplay of hydration and potential energy contributions, although the prediction of relatively compact structures suggest the importance of low potential energy contributions.

Like met–enkephalin, leu–enkephalin (H–Tyr–Gly–Gly–Phe–Leu–OH) is an endogenous pentapeptide in which the methionine residue has been replaced by a leucine residue. Qualitatively, the results for the hydrated forms of leu–enkephalin are similar to those for met–enkephalin [129].

Table 9: Comparison of hydration energies for met–enkephalin. The first column refers to the hydration model used in the function evaluations, which are performed at the global solutions for the hydration model listed in the second column. The total energy, E_{TOT}, is provided along with the contributions from hydration, E_{HYD}, nonbonded interactions (including hydrogen bonding), E_{NB}, electrostatic interactions, E_{ES}, and torsion, E_{TOR}. The last column provides the heavy atom root mean squared deviation between the global minimum energy structures of the hydration models listed in the first two columns.

	Global of	$\mathbf{E}_{TOT}$	$\mathbf{E}_{HYD}$	$\mathbf{E}_{NB}$	$\mathbf{E}_{ES}$	$\mathbf{E}_{TOR}$	(RMSD)
RRIGS	RRIGS	-50.01	-41.42	21.84	-31.46	1.02	0.00
	WE1	-47.87	-38.12	22.09	-32.61	0.78	2.83
	WE2	-47.91	-38.14	22.09	-32.63	0.76	2.83
	OONS	-47.17	-37.95	22.25	-32.13	0.66	2.66
	SCKS	-47.24	-35.61	21.47	-35.40	2.30	4.04
	JRF	-41.63	-46.69	23.29	-19.13	0.90	4.83
WE1	RRIGS	-26.60	-18.00	21.84	-31.46	1.02	2.83
	WE1	-30.31	-20.56	22.09	-32.61	0.78	0.00
	WE2	-30.31	-20.53	22.09	-32.63	0.76	0.01
	OONS	-29.01	-19.79	22.25	-32.13	0.66	0.80
	SCKS	-27.80	-16.17	21.47	-35.40	2.30	3.33
	JRF	-19.49	-24.55	23.29	-19.13	0.90	4.33
WE2	RRIGS	-29.87	-21.27	21.84	-31.46	1.02	2.83
	WE1	-33.26	-23.52	22.09	-32.61	0.78	0.01
	WE2	-33.27	-23.49	22.09	-32.63	0.76	0.00
	OONS	-32.01	-22.79	22.25	-32.13	0.66	0.80
	SCKS	-30.77	-19.15	21.47	-35.40	2.30	3.33
	JRF	-22.93	-27.99	23.29	-19.13	0.90	4.32

Table 10: Comparison of hydration energies for met–enkephalin. The first column refers to the hydration model used in the function evaluations, which are performed at the global solutions for the hydration model listed in the second column. The total energy, E_{TOT}, is provided along with the contributions from hydration, E_{HYD}, nonbonded interactions (including hydrogen bonding), E_{NB}, electrostatic interactions, E_{ES}, and torsion, E_{TOR}. The last column provides the heavy atom root mean squared deviation between the global minimum energy structures of the hydration models listed in the first two columns.

	Global of	$\mathbf{E}_{TOT}$	$\mathbf{E}_{HYD}$	$\mathbf{E}_{NB}$	$\mathbf{E}_{ES}$	$\mathbf{E}_{TOR}$	(RMSD)
OONS	RRIGS	-24.18	-15.59	21.84	-31.46	1.02	2.66
	WE1	-31.08	-21.33	22.09	-32.61	0.78	0.80
	WE2	-31.09	-21.31	22.09	-32.63	0.76	0.80
	OONS	-31.45	-22.23	22.25	-32.13	0.66	0.00
	SCKS	-29.57	-17.95	21.47	-35.40	2.30	3.38
	JRF	-19.60	-24.66	23.29	-19.13	0.90	4.12
SCKS	RRIGS	3.43	12.02	21.84	-31.46	1.02	4.04
	WE1	0.90	10.65	22.09	-32.61	0.78	3.33
	WE2	0.89	10.67	22.09	-32.63	0.76	3.33
	OONS	1.66	10.88	22.25	-32.13	0.66	3.38
	SCKS	-0.62	11.00	21.47	-35.40	2.30	0.00
	JRF	17.44	12.38	23.29	-19.13	0.90	3.78
JRF	RRIGS	-139.36	-130.77	21.84	-31.46	1.02	4.83
	WE1	-180.59	-170.84	22.09	-32.61	0.78	4.33
	WE2	-180.57	-170.79	22.09	-32.63	0.76	4.32
	OONS	-181.70	-172.48	22.25	-32.13	0.66	4.12
	SCKS	-171.67	-160.04	21.47	-35.40	2.30	3.78
	JRF	-283.76	-288.82	23.29	-19.13	0.90	0.00

5 Free energy modeling

Locating the global minimum *potential* energy or the global minimum *potential plus solvation* energy conformation is not sufficient because Anfinsen's thermodynamic hypothesis requires the minimization of the conformational free energy. Specifically, potential energy minimization neglects the entropic contributions to the stability of the molecule. An approximation to these entropic contributions can be developed by using information about low energy conformations. That is, once a sufficient ensemble of low energy minima has been identified, a statistical analysis can be used to estimate the relative entropic contributions, and thus the relative free energy, for conformations in the ensemble.

Therefore, the analysis of the free energy of peptides requires efficient methods for locating not only the global minimum energy structure but also large numbers of low energy conformers. A variety of methods have been used to find such stationary points on potential energy surfaces. For example, periodic quenching during a Monte Carlo or molecular dynamics trajectory can be used to identify local minima [222]. However, a drawback of these approaches is their inherent stochastic nature. In its original form, the αBB *deterministic* global optimization algorithm [3, 1, 4, 2, 9] has been shown to be an efficient method for finding the global minimum energy conformation for both unsolvated and solvated peptide systems [10, 128, 129]. Here, novel methods are proposed within the framework of the αBB algorithm to optimize the free energy of peptide systems. These modifications facilitate the generation of ensembles of low energy conformers, which can be used to identify the global minimum free energy conformation, as well as perform detailed free energy rankings.

In peptide systems, this entropic contribution arises from fluctuations around a local conformational state. There exists a number of procedures, including both exact and approximate calculations, that can be used to determine the entropic contributions, and thus the free energy, of peptide systems.

First, assume that the full conformational space R can be considered as the union of disjoint basins of attraction, and the conformational space associated with a given basin (denoted by γ) is defined by R_γ. The energy, E, is a function of the variable set θ, which corresponds to the set of dihedral angles used to describe the conformational state of the system. Each basin of attraction is characterized by a unique local minimum at position θ^*_γ, with a corresponding energy E^*_γ. That is, local minimization starting at any point in R_γ will lead to the local minimum at θ^*_γ. It should be noted that this approximation of the conformational space excludes all maxima and saddle point conformations.

For a given temperature, T, the probability that a peptide occupies the conformational space of a given basin (R_γ) can be described by a Gibbs/Boltzmann distribution :

$$p_\gamma = \frac{\int_{R_\gamma} exp(-\beta E(\theta))d\theta}{\int_R exp(-\beta E(\theta))d\theta} \tag{37}$$

Here β is equivalent to $1/k_B T$. If the numerator is redefined as the partition function (Z_γ) for the basin, Equation (37) can be rewritten as :

$$p_\gamma = \frac{Z_\gamma}{Z} \tag{38}$$

The total partition function for the entire conformational space is represented by Z. Since this function is described by a disjoint set of basins (R_γ), it is equivalent to the following form :

$$Z = \sum_\gamma Z_\gamma \tag{39}$$

Once the probability is known, the corresponding free energy, G_γ, associated with each basin can also be calculated :

$$G_\gamma = -\frac{\ln p_\gamma}{\beta} \tag{40}$$

Using these definitions, a rigorous procedure can be envisioned for calculating the exact probability associated with a given basin. First, a sample of conformations must be generated with initial starting energies E_i, as defined by the total set I. Each structure is minimized to identify its corresponding basin minimum (θ^*_γ). These structures define the set $I(\gamma)$ (i.e., those structures associated with basin γ). As the sampling goes to infinity, the probability associated with basin γ can be calculated by the following expression :

$$p_\gamma^{exact} = \frac{\sum_{i(\gamma)\in I(\gamma)} exp(-\beta E_{i(\gamma)})}{\sum_{i\in I} exp(-\beta E_i)} \tag{41}$$

Obviously, such a method is intractable for large systems, and this is the impetus for developing approximate methods.

6 Harmonic approximation

A tractable method for including entropic effects for proteins relies on the concept of the harmonic approximation. Initially, the theoretical development of this approximation for polymer systems generated debate in the literature [82, 83, 69]. In the work of [82] [82] a classical rigid model was used to characterize a partition function based on the fixed bond length and bond angle assumptions. In contrast, [69] [69] derived a different partition function using a classical flexible model. Later analysis by [83] [83] actually showed that the flexible model was also applicable to the fixed bond length and bond angle system (i.e., a peptide described by the internal coordinate system).

In either case (i.e., rigid or flexible), entropic contributions can be calculated by employing an harmonic approximation [83]. The fundamental concept is to characterize the basin of attraction (γ) by the properties of its corresponding local minimum (θ^*_γ), and not by a random sampling of conformations. These properties include the local minimum energy value, E^*_γ, and the convexity around the local minimum. Essentially, the convexity measure is used to approximate the basin of attraction region as a hyperparabola centered at the local minimum. Therefore, the anharmonic nature of the true basin, which defines the deviation from approximated harmonic behavior, controls the error associated with this assumption.

At each minimum (θ_γ^*) the harmonic approximation to the entropy can be evaluated using the following expression :

$$S_\gamma^{approx} = -\frac{k_B}{2} \ln \left[Det \left(H_\gamma \right) \right] + \hat{f}(T) \tag{42}$$

Here $Det\,(H_\gamma)$ refers to the determinant of the Hessian (second derivative matrix) evaluated at the local minimum θ_γ^*. The function $\hat{f}(T)$ is an additive term that is only dependent on temperature. The approximated free energy can then be calculated by combining the energetic and entropic contributions through the follow expression :

$$G_\gamma^{approx} = E_\gamma^* - TS_\gamma^{approx} + \bar{f}(T) \tag{43}$$

By substituting the harmonic entropic approximation from Equation (42), Equation (43) becomes :

$$G_\gamma^{approx} = E_\gamma^* + \frac{1}{2\beta} \ln \left[Det \left(H_\gamma \right) \right] + \tilde{f}(T) \tag{44}$$

In this equation, it becomes evident that the free energy for a given basin is estimated using only the properties of the corresponding local minimum, that is, the local minimum energy (E_γ^*) and a measure of local convexity ($Det\,(H_\gamma)$). A temperature dependent term, $\tilde{f}(T)$, is included, although it does not affect relative free energy comparisons.

Expressions for the probabilities and partition functions can also be developed. By combining Equations (38), (40) and (44), an approximation for the partition function of a given basin can be written as :

$$\ln Z_\gamma^{approx} = -\beta E_\gamma^* - \frac{\ln \left[Det \left(H_\gamma \right) \right]}{2} - \beta \tilde{f}(T) + \ln Z \tag{45}$$

A further simplification can be made by realizing that $-\beta\tilde{f}(T)$ and $\ln Z$ are constant for a given temperature (i.e., $f(T) = -\beta\tilde{f}(T) + \ln Z$). Equation (45) can be rewritten as :

$$Z_\gamma^{approx} = \left[\frac{1}{\left[Det \left(H_\gamma \right) \right]} \right]^{1/2} \exp\left(-\beta E_\gamma^* \right) f(T) \tag{46}$$

Finally, by using Equation (39), an approximate probability associated with a given basin (γ) can be calculated using the following equation :

$$p_\gamma^{approx} = \frac{\left[\left[Det \left(H_\gamma \right) \right] \right]^{-1/2} \exp\left(-\beta E_\gamma^* \right)}{\sum_{i=1}^{N} \left[\left[Det \left(H_i \right) \right] \right]^{-1/2} \exp\left(-\beta E_i^* \right)} \tag{47}$$

As expected, the $f(T)$ term disappears, and the statistical weight becomes a function of only the temperature (through β), the local minimum energy value and the measure of convexity. In order to develop a meaningful comparison of relative free energies, the total partition function (i.e., the denominator of Equation (47)) must include an adequate ensemble of low energy local minima, as well as the global minimum energy conformation.

These probabilities can be used to estimate the occupancy of each individual basin, or summed in order to calculate cumulative probabilities for an ensemble of structures exhibiting similar physical or energetic properties. It should be noted that the determination of free energy using the harmonic approximation does not require the explicit inclusion of a contribution based on the density of states. That is, the harmonic approximation decomposes the energetic states within a basin of attraction into one energetic value represented by the local minimizer of the basin. In contrast to counting methods, which estimate probabilities based on the density of states, the contribution of each structure should be accounted for only once. Therefore, using the harmonic approximation requires a structural comparison of all local minimizers.

The probabilities obtained through the harmonic approximation can also be used to calculate thermodynamic quantities. Once the set of unique minimizers has been identified, these structures can be ranked according to their free energy values, and then divided into bins of a specified energy width. Probabilities for each bin can be calculated by summing the individual probabilities (as defined in Equation (47)) :

$$P_j^{approx} = \sum_{\gamma=1}^{n_j} p_\gamma^{approx} \tag{48}$$

Here P_j^{approx} signifies the probability for energy bin j. The summation includes the n_j individual probabilities (p_γ^{approx}) belonging to bin j. Average thermodynamic quantities can now be estimated using the equations with the following form :

$$< E >_T = \sum_j P_j^{approx} < E >_j \tag{49}$$

Here the total average energy, $< E >_T$, is calculated by summing the bin probabilities multiplied by the mean energy of bin j, $< E >_j$.

7 Free energy problem formulation

As before, the energy minimization problem for proteins is formulated as a nonconvex nonlinear optimization problem. The inclusion of free energy modeling into the protein folding problem does not change the general formulation. However, an additional condition must be satisfied; that is, an ensemble of local minimum low energy conformations must be generated along with the global minimum energy conformation. Once this ensemble has been compiled, a free energy ranking can be performed using the harmonic approximation presented in the previous section.

Several rigorous methods can be envisioned for locating local minimum energy conformations using the αBB deterministic global optimization approach. As an introduction to the ideas used here, two rigorous approaches for finding all local minimum energy conformations are discussed.

The first method relies on the introduction of a single inequality constraint to the problem formulation given by (34). The new formulation is :

$$\min \quad E(\phi_i, \psi_i, \omega_i, \chi_i^k, \phi_j^N, \phi_j^C) \tag{50}$$

$$\begin{aligned} \text{subject to} \quad & (E^* - E) + \epsilon^* < 0 \\ -\pi \leq \; & \phi_i \leq \pi, \quad i = 1, \ldots, N_{RES} \end{aligned}$$

$$
\begin{aligned}
-\pi &\leq \psi_i \leq \pi, \;\; i = 1, \ldots, N_{RES} \\
-\pi &\leq \omega_i \leq \pi, \;\; i = 1, \ldots, N_{RES} \\
-\pi &\leq \chi_i^k \leq \pi, \;\; i = 1, \ldots, N_{RES}, \; k = 1, \ldots, K^i \\
-\pi &\leq \phi_j^N \leq \pi, \;\; j = 1, \ldots, J^N \\
-\pi &\leq \phi_j^C \leq \pi, \;\; j = 1, \ldots, J^C
\end{aligned}
$$

The additional constraint requires that the objective function values be larger than the energy value at some local (or global) minimum, as denoted by E^*, plus a positive parameter, ϵ^*. When $\epsilon^* = 0$, the solution of the corresponding global optimization problem will give the best local minimum energy conformation with an energy larger than E^*. The original formulation given by (34) is actually a special case of this problem in which $E^* = -\infty$ and $\epsilon^* = 0$. That is, in (34) no bounds are placed on the value of the objective function, E. The global minimum energy conformation is only required to take some finite value. In order to locate all local minima, a set of global optimization problems must be solved iteratively with updating of the parameter E^*.

The problem of finding all local minimum energy conformations can also be formulated as a single global optimization problem, which can be deterministically solved using the αBB algorithm [159]. This method stems from the idea that all stationary points (i.e., minima, maxima and transition states) of the energy hypersurface satisfy the constraint $\nabla E(\theta) = \mathbf{0}$. This can be written as :

$$
\frac{\partial E(\theta)}{\partial \theta_i} = 0 \;, \qquad i = 1, \ldots, N_\theta \tag{51}
$$

Here N_θ represents the total number of dihedral angles defined by the variable set θ. The problem of finding local minima is equivalent to finding all solutions of Equation (51) for which the Hessian of E is positive definite.

The problem posed in Equation (51) involves the solution of a system of nonlinear equations. The identification of all multiple global solutions requires the use of a deterministic global optimization method, as outlined in Section II.B. The application of this method to protein system will be described fully in Section IV.B.

Both methods for rigorously locating all local minimum energy conformations have some disadvantages. On one hand, the first approach should effectively locate low energy conformers in order of increasing energy. However, locating each minimum requires the solution of a full global optimization problem. The second approach avoids this drawback because it can be solved as a single global optimization problem. However, when dealing with a high dimensional search space, the number of necessary subdivisions may be computationally inhibitive. In addition, this method will potentially locate stationary points other than local minima. Therefore, the development of other methods for locating low energy local minimum energy conformations were pursued.

8 Ensemble of local minimum energy conformations

Since the number of local minima on a given energy hypersurface may become astronomically large (e.g., the number of local minima for met-enkephalin is estimated to be on the order of 10^{11}

[150]), methods that do not necessarily provide all local minima were developed. Specifically, it was determined that the generation of ensembles of low energy conformers is possible through algorithmic modifications of the general αBB procedure. Rigorous implementation of the global optimization algorithm requires the minimization of a *convex* lower bounding function in each domain. The unique solution θ for each lower bounding minimum can then used as a starting point for the minimization (or function evaluation) of the original energy function in the current domain. In the case of local minimization, each partitioned region provides a single minimum energy conformation as the algorithm proceeds. Using this information, along with the global minimum energy conformation, a list of low energy conformers can be constructed.

A method for increasing the number of local minima produced within each subdomain would involve the selection of multiple random starting points for minimizing the upper bounding function. At first, this approach appears to be equivalent to choosing random points for local minimization. Initially, when the subdomains constitute significant portions of the original domain space, this is the case. However, as the separation between lower and upper bounds decreases, the subdomains are localized in regions of low energy. Therefore, the random point selection is localized in regions which contain low energy local minima.

However, this approach does not take advantage of the information provided by the lower bounding functions. Rigorously, these functions posses a single minimum in each subdomain. Since the choice of α affects the convexity of the lower bounding functions, the α values can be modified to ensure a certain nonconvexity in these functions. In this case, the lower bounding functions possess multiple minima, and these functions can be minimized several times in each domain. In addition, since the lower bounding functions smooth the original energy hypersurface, the location of these multiple minima provide information on the location of low energy minima for the upper bounding function. Therefore, by using the location of the minima of the lower bounding function as starting points for local minimization of the upper bounding function, an improved set of low energy conformations can be identified. As before, these conformations are also localized in those domains with low energy as the subdomains decrease in size. This energy directed approach (EDA) is represented schematically in Figure 17.

The basic steps of the algorithm, which are qualitatively similar to those outlined in Figure 12, are as follows:

(1) The initial best upper bound is set to an arbitrarily large value. The original domain is partitioned along one of the global variables. α values are initially chosen to be constant ($\alpha = \alpha_0$) for all global variables.

(2) The lower bounding function (L) is constructed in each hyper–rectangle. Three local minimization are performed using the following procedure :

- **(A)** 50 random points are generated and used for function evaluations.
- **(B)** The point with the minimum value is used as a starting point for local minimization of L using NPSOL, with calls (through PACK) to ECEPP/3 and possibly the RRIGS solvation module.

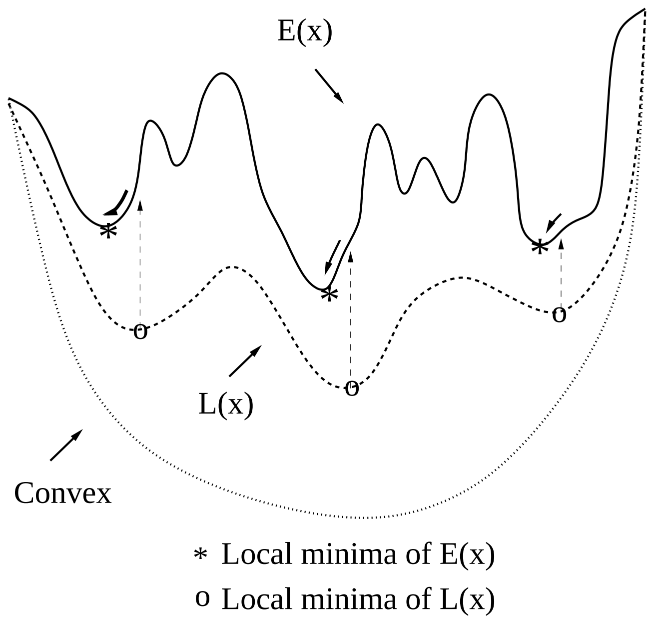

Figure 17: Using multiple lower bound minima to find low energy conformers of the upper bounding function.

(C) The unique solutions are stored

If the minimum valued solution (of all local minima of L in this subdomain) is greater than the current best upper bound the subdomain is fathomed.

(3) The unique local minima (points) for L are used as initial starting points for local minimizations of the upper bounding function (E) in each hyper–rectangle. Again, the appropriate calls are made to PACK and the potential and solvation energy modules. Two additional minimizations are performed using the following procedure :

(A) 50 random points are generated and used for function evaluations.

(B) The point with the minimum value is used as a starting point for local minimization of E using NPSOL, with calls (through PACK) to ECEPP/3 and possibly the RRIGS solvation module.

In all cases, the UBC (upper bound check) module is also called. UBC checks that the absolute value of each gradient in the objective function gradient vector is below a specified tolerance (10^{-6} kcal/mol/deg). If a gradient does not satisfy this check the corresponding variable bounds are incrementally increased and the problem is solved with the previous point used as the initial starting point. This process is repeated until the gradient constraints are satisfied or an iteration limit is exceeded. UBC also employs algorithms to calculate the second derivative matrix [183], which is used to verify that the upper bound solution is a local minimum; that is, the Hessian matrix must be positive semi–definite. If the matrix is not positive semi–definite or the gradient checks are not satisfied, the upper bound solution is rejected. All local minima are stored.

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

(5) The hyper–rectangle with the current minimum value for L (this is the minimum value of all local minima of L in each subdomain) is selected and partitioned along one of the global variables. All α values are updated according to the following rule:

$$\alpha = \alpha_0 R^L \tag{52}$$

In this equation α_0 refer to the initial values from Step 1. R is a reduction parameter $(0 < R \leq 1)$, and L refers to the current level in the branch and bound tree. For $R = 1$ the α values are kept constant at the initial value, α_0.

(6) If the best upper and lower bounds are within the ϵ tolerance, or a maximum iteration limit has been exceeded, the program will terminate, otherwise it will return to Step 2.

A second approach incorporates free energy information into the branch and bound algorithm. Specifically, harmonic entropic contributions are calculated and included at each minima of the upper and lower bounding functions. In this way, the progression of lower and upper bounds includes a temperature dependent entropic term. A similar modification to the Monte Carlo minimization method has also been proposed [234], and has been shown to be effective in locating low energy conformers of peptides [165, 166].

The problem formulation is identical to the one given in (34). That is, the minimization of E and L are still performed using only potential and solvation energy contributions. However, once local minima have been located, the free energy is calculated by the following expression:

$$G = U_{Min} + \frac{1}{2\beta} \ln \left[Det\left(H_{Min}\right) \right] \tag{53}$$

This equation is similar to Equation (44), although the additive term $f(T)$ has been omitted because it is a function of temperature only. U_{Min} represents the local minimum energy of E or L, and $Det\left(H_{Min}\right)$ is the determinant of the Hessian evaluated at this local minimum. The specification of a thermodynamic temperature $(\beta = 1/k_B T)$ is required as an additional input parameter.

A single rigorous application of the αBB algorithm to this problem will result in the identification of the global minimum free energy at a given temperature. However, the goal is to identify an ensemble of low energy and, in this case, low free energy conformers so that a free energy ranking and comparison can be made. Therefore, the algorithmic steps for the Free Energy Directed Approach (FEDA) are similar to those for EDA, with the additional evaluation of the free energy (G) at each local minima of E and L. The thermodynamic temperature used in Equation (53) must be specified as an additional input parameter.

9 Free energy computational studies

The EDA was first applied to the isolated form of met-enkephalin. Met–enkephalin (H–Tyr–Gly–Gly–Phe–Met–OH) is an endogenous opioid pentapeptide found in the human brain, pituitary, and peripheral tissues and is involved in a variety of physiological processes. The peptide consists of 24 independent torsional angles and a total of 75 atoms, and has played the role of a benchmark molecular conformation problem. All 24 dihedral angles were considered variable, with the 10 dihedral angles of the backbone residues acting as global variables (variables on which branching occurs). For both peptides, the EDA algorithm detailed above, was applied 10 times. The input conditions correspond to initial α values of 5 and 10, with a subsequent reduction of these values based on the current level in the branch and bound tree.

Once the ensemble of local minima had been compiled, a set of distinct conformations was identified by checking for repeated and symmetric conformations. In addition, a conformation was only considered unique if at least one dihedral angle differed by at least 50^o when comparing each pair of conformations. These conformations were then used to generate results and distributions according to energy and free energy values. Energy bins were used to characterize a group of distinct structures between a range of energy values (every 0.5 kcal/mol) relative to the global minimum energy structure. For example, Bin 1 contains structures that are 0.0-0.5 kcal/mol above the global minimum energy structure, Bin 2 contains structures that are 0.5-1.0 kcal/mol above the global minimum energy structure, etc.

In the case of isolated met-enkephalin, the 10 (EDA) runs generated a total of 83908 distinct local minima. The potential energy global minimum (PEGM) conformation for met-enkephalin possesses an energy of -11.707 kcal/mol. This conformation exhibits a type II' β–bend along the N–C' peptidic bond of Gly^3 and Phe^4. Essentially, this structure corresponds to the free energy global minimum (FEGM) conformation for a temperature of 0 K, that is, when entropic contributions are not included. When considering the harmonic free energy, the prediction of the FEGM can be calculated over a range of temperatures. Table 11 provides information on the FEGM for temperatures ranging from 100 K to 500 K.

As Table 11 shows, the PEGM persists as the FEGM at a temperature of 100 K. However, at the next three temperature points (i.e., 200 K, 300 K, 400 K) the FEGM exhibits a potential energy contribution 1.808 kcal/mol higher than the PEGM. The ϕ and ψ values for this structure are also significantly different than those for the PEGM. In fact, the conformational code (B*AAAE) indicates that the central residues display an α helical configuration. At a temperature of 500 K, the FEGM structure changes again, while the potential energy difference between the FEGM and PEGM increases to 5.369 kcal/mol. These differences suggest that the inclusion of entropic contributions greatly affects the relative stability of individual low energy structures. In addition, as the temperature increases, the stability offered by entropic contributions offsets substantial differences in potential energy.

Table 12 provides information on the distribution of distinct low free energy minima within 8.0 kcal/mol of the FEGM for a range of temperatures. For a given temperature the general trend indicates a large increase in the number of minima as the free energy increases above the FEGM.

Table 11: Dihedral angle values for PEGM and FEGM structures of isolated met-enkephalin using EDA. The temperatures are provided in the first row. The last two rows indicate the harmonic free energy (kcal/mol) and the potential energy value (kcal/mol), respectively.

Residue	DA	PEGM	100 K	200 K	300 K	400 K	500 K
Tyr_1	ϕ	-83.4	-83.4	179.8	179.8	179.8	90.2
	ψ	155.8	155.8	-18.2	-18.2	-18.2	149.1
	ω	-177.1	-177.1	-178.1	-178.1	-178.1	177.5
	χ_1	-173.2	-173.2	178.2	178.2	178.2	169.8
	χ_2	79.3	79.3	81.3	81.3	81.3	-108.2
	χ_3	-166.3	-166.3	177.3	177.3	177.3	177.6
Gly_2	ϕ	-154.3	-154.3	-59.8	-59.8	-59.8	-66.1
	ψ	85.8	85.8	-37.6	-37.6	-37.6	87.5
	ω	168.5	168.5	-178.8	-178.8	-178.8	-173.4
Gly_3	ϕ	83.0	83.0	-67.0	-67.0	-67.0	147.2
	ψ	-75.0	-75.0	-40.1	-40.1	-40.1	-36.7
	ω	-170.0	-170.0	179.7	179.7	179.7	175.1
Phe_4	ϕ	-136.9	-136.9	-70.9	-70.9	-70.9	-92.5
	ψ	19.1	19.1	-39.5	-39.5	-39.5	-34.7
	ω	-174.1	-174.1	-179.8	-179.8	-179.8	-179.1
	χ_1	58.9	58.9	173.9	173.9	173.9	179.1
	χ_2	94.5	94.5	-102.6	-102.6	-102.6	74.7
Met_5	ϕ	-163.5	-163.5	-161.0	-161.0	-161.0	-154.7
	ψ	160.9	160.9	122.1	122.1	122.1	135.3
	ω	-179.8	-179.8	-178.0	-178.0	-178.0	179.9
	χ_1	52.9	52.9	-174.7	-174.7	-174.7	-172.6
	χ_2	175.3	175.3	174.0	174.0	174.0	175.1
	χ_3	-179.9	-179.9	179.0	179.0	179.0	179.9
	χ_4	-178.6	-178.6	-60.1	-60.1	-60.1	-60.0
G		-11.707	-2.499	6.151	14.175	22.200	29.592
E		-11.707	-11.707	-9.899	-9.899	-9.899	-6.338

Table 12: Number of distinct minima in bins for isolated met-enkephalin using EDA. Each bin represents a 0.5 kcal/mol range above the previous bin. The temperatures are given in the first row.

Bin	0 K	50 K	100 K	150 K	200 K	250 K	300 K	350 K	400 K	450 K	500 K
1	2	1	2	10	6	3	3	4	16	16	8
2	3	5	13	22	12	9	15	24	18	21	31
3	12	25	36	58	52	42	40	40	59	69	77
4	45	48	55	105	105	100	101	115	164	184	184
5	49	69	120	233	199	206	213	249	309	397	475
6	90	125	263	451	435	403	410	491	726	893	918
7	166	292	467	806	763	765	848	1043	1438	1655	1687
8	303	497	766	1250	1297	1362	1524	1906	2464	2821	2695
9	552	776	1233	1929	2079	2247	2601	3069	3932	4284	4111
10	840	1177	1710	2915	3168	3475	3927	4707	5774	6030	5562
11	1121	1675	2681	3879	4355	4899	5708	6655	7573	7775	7116
12	1618	2467	3526	5303	5935	6572	7364	8333	9437	9448	8721
13	2331	3223	4491	6821	7619	8360	9203	10228	10730	10473	9719
14	2973	4050	6037	8058	8834	9712	10598	11244	11651	11285	10630
15	3747	5250	7258	9031	9821	10585	11504	11939	11915	11396	10745
16	4588	6422	8053	8587	9687	10958	11563	11432	9406	8482	8338

Several exceptions to this trend occur at high temperature and large bin number. In these cases, the number of minima remains constant or even decreases slightly. This is most likely due to an inadequate sampling of higher potential energy minima. For a given bin, it is also apparent that the clustering of low free energy structures increases with temperature. This increased density of the free energy bins indicates that increases in energy are offset by entropic contributions.

These observations are also supported by the information shown in Figure 18. This plot displays the range of potential energy in free energy bins at temperatures of 250 and 500 K, with the potential energy bins included for comparison. As expected, the potential energy values for the free energy bins increase with increasing temperature. In addition, the range of potential energy values increases in higher free energy bins. It is interesting to note that this occurs because the minimum potential energy is relatively (i.e., within a few kcal/mol of the PEGM) low for each bin, whereas the maximum potential energy value increases in higher bins. The corresponding differences are also greater at higher temperature. For example, at 500 K some bins exhibit a 20 kcal/mol range in potential energy. These trends explain the increased number of low free energy conformers. That is, bins of low free energy contain conformers of relatively high potential energy because of their more stabilizing entropic contributions. The plot also implies that the PEGM appears in bins 3 and 10 for temperatures of 250 and 500 K, respectively.

Relative free energies were also calculated for clusters of low energy conformers. This analysis is useful because it is difficult to capture the true accessibility of individual structures based on a point-wise approximation of entropic effects. That is, the harmonic free energy approximation

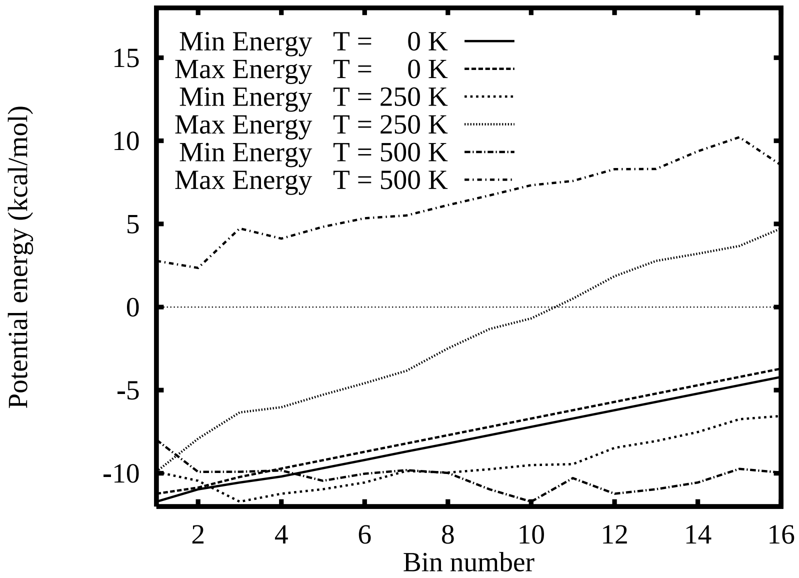

Figure 18: Potential energy comparison for isolated met-enkephalin using EDA. Minimum and maximum potential energies versus bin number are plotted for three temperatures : T = 0 K, 250 K, 500 K.

does not provide a continuous free energy landscape. By clustering structures into larger groups, it is hoped that the error associated with these estimates will average out. Typically, structures are clustered by calculating and comparing root mean squared deviations. Since the enkephalin peptide is relatively small, structures were grouped based on the Zimmerman codes for the central residues of the peptide [253]. Specifically, for met-enkephalin, structures were said to belong to the same cluster if the central 3 residues possessed the same 3 code letters based on the Zimmerman classification [253]. The relative free energy of a cluster was calculated by the following equation :

$$G_{cluster} = -\frac{\ln \sum_{i \in C} p_i^{approx}}{\beta} \tag{54}$$

In Equation (54) the individual p_i^{approx}, which refers to the statistical weight based on the harmonic approximation, are summed for the set of conformations belonging to a particular cluster (C). These individual probabilities were calculated by normalizing each probability with respect to the overall probability at a given temperature :

$$p_i^{approx} = \frac{\exp[-\beta(G_o^{approx} - G_i^{approx})]}{\sum_j \exp[-\beta(G_o^{approx} - G_j^{approx})]} \tag{55}$$

A reference free energy, G_o^{approx}, was used to normalize the probabilities at each temperature point. All free energies, G_o^{approx}, G_i^{approx} and G_j^{approx}, refer to the harmonic approximation of the free energy as calculated using Equation (44). The denominator, which represents the total probability at a given temperature, is calculated by summing over the set of all conformers.

Table 13: Clustered relative free energies for isolated met-enkephalin using the EDA. From left to right, the information provided in this table includes : temperature, Zimmerman code^i, number of individual structures in cluster, total probability ($\sum_i p_i^{approx}$) and free energy of cluster ($G_{cluster}$).

Temp (K)	Code	Number	$\sum_i p_i^{approx}$	$G_{cluster}$
	DC*B	113	0.636	0.0899
100	CC*B	136	0.0794	0.503
	C*DE	557	0.0765	0.511
	AAA	323	0.230	0.585
200	DC*A	1828	0.213	0.615
	C*DE	676	0.192	0.656
	CD*A	2685	0.297	0.723
300	DC*A	1843	0.100	1.372
	AAA	328	0.0990	1.379
	CD*A	2654	0.219	1.209
400	DC*A	1799	0.0452	2.461
	AAA	329	0.0380	2.600
	CD*A	2449	0.112	2.174
500	C*C*A	1361	0.0256	3.640
	C*AE	1463	0.0229	3.752

The relative free energies for clusters of met-enkephalin structures are given in Table 13. At each temperature point the Zimmerman code and corresponding data for the top 3 clusters are listed. The results indicate that the structure exhibiting the individual lowest free energy does not always belong to the cluster with lowest free energy. At 100 and 200 K the DC*B and AAA clusters are consistent with the structures of the FEGM. However, although the FEGM retains the AAA structure at 300 and 400 K, the group of structures possessing the lowest $G_{cluster}$ at these temperatures exhibits a CD*A Zimmerman code. This is, at least in part, attributable to the large number of structures grouped in this cluster. In contrast to the α-helical type structure for the FEGM, the CD*A structures possess elements of a β-turn conformation. Specifically, the lowest free energy conformer exhibiting a CD*A structure at 300 and 400 K, possesses a type II β-bend along the Gly^2–Gly^3 backbone.

FEDA was also applied to the isolated form of met-enkephalin. For this approach, the thermodynamic temperature appears as an input parameter, and these values had to be specified along with initial α values. Several methods can be envisioned for initializing the FEDA. For example, if the goal is to characterize the low free energy conformers at a single temperature, a full set of FEDA runs could be performed for that temperature. This type of search should efficiently locate the global and many low free energy conformers for that temperature. However, the goal was to effectively characterize the FEGM and low free energy conformers over a range of temperatures. Therefore each of the 10 (FEDA) runs were conducted at a unique temperature point in the range of 50 to 500 K. The details of the conditions for these runs are given in Table 14.

In total, 87974 distinct local minima were found after compiling the results from the 10 (FEDA)

Table 14: Input parameters used for FEDA runs. α_0 refers to the initial α values used for all global variables. R refers to the reduction rate applied at each level of the branch and bound tree. T refers to the thermodynamic temperature at which the free energy was calculated.

Run No.	α_0	R	T (K)	Run No.	α_0	R	T(K)
1	5	0.90	50	6	5	0.90	300
2	5	0.90	100	7	5	0.90	350
3	5	0.90	150	8	5	0.90	400
4	5	0.90	200	9	5	0.90	450
5	5	0.90	250	10	5	0.90	500

runs for isolated met-enkephalin. The PEGM and FEGM found using the FEDA are displayed in Table 15. It should be noted that when comparing PEGM for the EDA and FEDA, both structures possess the same potential energies, but a different set of dihedral angles. However, these structures are actually the same. That is, the different values of χ_2 and χ_3 for Tyr_1 represent a degenerate state for tyrosine, which is generated by rotating both of these dihedral angles by 180^o. An important observation is that at 200 K the FEDA method predicts a slightly lower FEGM. The structure possesses a lower potential energy (-10.547 vs. -9.899 kcal/mol) and exhibits a free energy value that is 0.044 kcal/mol lower than the EDA predicted FEGM. The remaining FEGM predictions are consistent for the two approaches.

An analysis of the distribution of distinct minima, as given by Table 16, reveals that the results are qualitatively consistent with those produced by the EDA. It should be noted that in all cases, the lowest free energy bin is as populated as the corresponding EDA bins, which indicates that each run using the FEDA was able to find a better distribution of low free energy conformers near the FEGM. This is not unexpected, considering that the FEDA runs were conducted at the same discrete temperature points used in the analysis. However, when comparing the populations of higher energy bins at low temperatures, the number of minima are larger for the EDA. Some of this variation, especially near the 150 to 200 K range is probably due to the lower FEGM found by the FEDA. In general, the FEDA seems to provide a denser distribution of distinct minima at higher temperatures and large bin number.

A comparison of the relative efficiencies for the EDA and FEDA to generate low energy local minima can also be made by examining Figure 19. In this plot the cumulative fraction of conformers, which is equal to the total number of unique conformers within the first 8, 12 and 16 energy bins over the total number of unique conformers, is given as a function of temperature. It is apparent that both approaches are highly efficient. For example, at 400 K approximately 90 percent of the total unique conformations identified are in the top 16 free energy bins, which ranges up to 8 kcal/mol above the FEGM. The lower fractions at lower temperatures indicate that a relatively large number of conformations have high potential energies, and that these energetic differences are not offset by entropic effects at low temperatures. A more subtle comparison can be made by observing that the EDA cumulative fractions are generally higher for temperatures lower than 400

Table 15: Dihedral angle values for PEGM and FEGM structures of isolated met-enkephalin using FEDA. The temperatures are provided in the first row. The last two rows indicate the harmonic free energy (kcal/mol) and the potential energy value (kcal/mol), respectively.

Residue	DA	PEGM	100 K	200 K	300 K	400 K	500 K
Tyr_1	ϕ	-83.4	-83.4	-163.1	179.8	179.8	-90.2
	ψ	155.8	155.8	-40.5	-18.2	-18.2	149.1
	ω	-177.1	-177.1	-177.7	-178.1	-178.1	177.5
	χ_1	-173.2	-173.2	-172.2	178.2	178.2	169.8
	χ_2	-100.7	-100.7	93.2	81.3	81.3	71.8
	χ_3	13.7	13.7	-177.2	177.3	177.3	-2.4
Gly_2	ϕ	-154.3	-154.3	65.1	-59.8	-59.8	-66.1
	ψ	85.8	85.8	-89.7	-37.6	-37.6	87.5
	ω	168.5	168.5	174.1	-178.8	-178.8	-173.4
Gly_3	ϕ	83.0	83.0	-152.6	-67.0	-67.0	147.2
	ψ	-75.0	-75.0	34.4	-40.1	-40.1	-36.7
	ω	-170.0	-170.0	-178.9	179.7	179.7	175.1
Phe_4	ϕ	-136.8	-136.8	-155.4	-70.9	-70.9	-92.5
	ψ	19.1	19.1	159.8	-39.5	-39.5	-34.7
	ω	-174.1	-174.1	179.2	-179.8	-179.8	-179.1
	χ_1	58.9	58.9	52.1	173.9	173.9	179.1
	χ_2	-85.5	-85.5	82.9	-102.6	-102.6	74.7
Met_5	ϕ	-163.5	-163.5	-79.3	-161.0	-161.0	-154.7
	ψ	160.9	160.9	130.4	122.1	122.1	135.3
	ω	-179.8	-179.8	-178.7	-178.0	-178.0	179.9
	χ_1	52.9	52.9	-66.8	-174.7	-174.7	-172.6
	χ_2	175.3	175.3	179.8	174.0	174.0	175.1
	χ_3	-179.9	-179.9	-179.9	179.0	179.0	179.9
	χ_4	-178.6	-178.6	-60.0	-60.1	-60.1	180.0
G		-11.707	-2.499	6.107	14.175	22.200	29.592
E		-11.707	-11.707	-10.547	-9.899	-9.899	-6.338

Table 16: Number of distinct minima in bins for isolated met-enkephalin using FEDA. Each bin represents a 0.5 kcal/mol range above the previous bin. The temperatures are given in the first row.

Bin	0 K	50 K	100 K	150 K	200 K	250 K	300 K	350 K	400 K	450 K	500 K
1	2	1	3	10	8	5	5	6	17	15	8
2	3	6	14	9	10	11	16	23	19	23	30
3	12	26	38	52	53	43	42	41	56	63	86
4	46	48	55	87	91	100	97	107	156	188	193
5	47	69	116	180	189	205	208	249	324	407	478
6	87	122	259	373	400	391	403	481	721	898	988
7	161	290	470	654	730	758	846	1051	1476	1801	1756
8	297	488	760	1063	1246	1368	1524	1936	2576	2966	3052
9	543	762	1182	1637	1918	2188	2597	3181	4136	4618	4538
10	828	1140	1624	2413	2996	3511	4032	4863	6033	6481	6070
11	1066	1560	2569	3542	4193	4852	5726	6791	8047	8466	7832
12	1527	2404	3433	4735	5785	6616	7499	8630	9989	10069	9426
13	2244	3070	4470	6288	7382	8341	9315	10632	11286	11130	10484
14	2818	4004	5833	7451	8649	9727	10862	11833	12430	11937	11102
15	3657	5064	7075	8723	9617	10818	12004	12606	12358	11968	11238
16	4472	6257	7848	8718	10108	11295	12167	12003	9952	8640	8576

K. Although the total number of unique conformations is slightly lower for the EDA, this trend indicates that the EDA is more efficient at filling low energy bins, especially at lower temperatures.

The results for the cluster analysis of the FEDA met-enkephalin structures are given in Table 17. There are some differences between the EDA and FEDA cluster free energies, although the overall trend is the same. At all temperatures, excluding 200 K, the cluster exhibiting the lowest cluster free energy is the same as in the EDA analysis. At 200 K, the FEDA predicts the AAA cluster as having a slightly higher free energy than the C*DE cluster , which only appears as the third cluster in Table 13. In both analyses, the transition from the ground state DC*B cluster to the CD*A cluster as temperature increases, is evident.

Since both the EDA and FEDA provide large amounts of statistical information for the peptide system, this data was used to perform a simple thermodynamic analysis of the folding process. It is widely accepted that the folding of peptides progresses successively. The first step of this process is typically associated with a structural collapse, that is, a transition from random extended structures to an ensemble of compact structures. This transition should also be associated by significant changes in the description of the ensemble as temperature changes. For example, a peak in the specific heat at the transition temperature indicates a steep decrease in average potential energy of the ensemble. In order to verify that such a transition occurs for met-enkephalin, the specific heat was calculated using the following expression :

$$C = \frac{\beta^2 \left(< E^2 >_T - < E >_T^2\right)}{N} \tag{56}$$

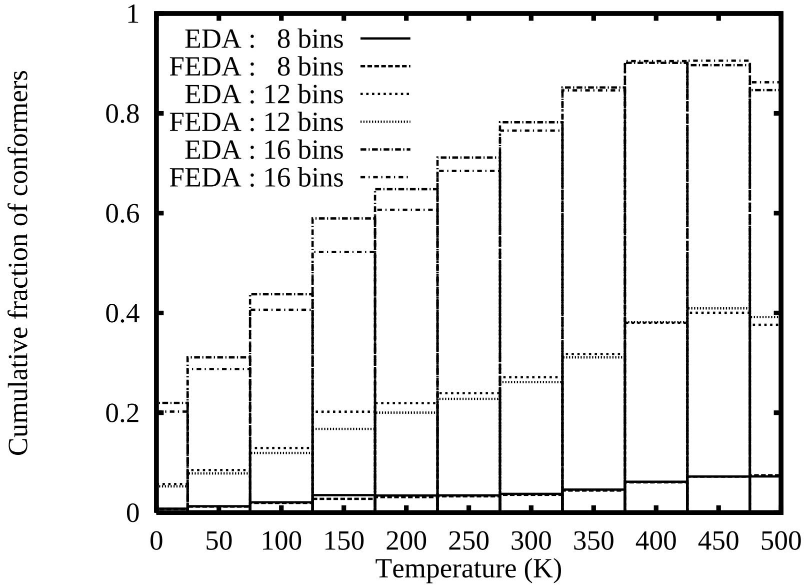

Figure 19: Plot of cumulative fraction of low energy conformers for isolated met-enkephalin, which is equal to the number of unique conformers within the first 8, 12 and 16 energy bins over the total number unique conformers, versus temperature. Both EDA and FEDA data are plotted.

Table 17: Clustered relative free energies for isolated met-enkephalin using the FEDA. From left to right, the information provided in this table includes : temperature, Zimmerman code[i], number of individual structures in cluster, total probability ($\sum_i p_i^{approx}$) and free energy of cluster ($G_{cluster}$).

Temp (K)	Code	Number	$\sum_i p_i^{approx}$	$G_{cluster}$
	DC*B	107	0.532	0.125
100	C*DE	990	0.232	0.291
	CC*A	1604	0.0636	0.547
	C*DE	1275	0.331	0.439
200	AAA	322	0.209	0.623
	DC*A	1729	0.174	0.694
	CD*A	2128	0.263	0.796
300	C*DE	1360	0.125	1.239
	AAA	327	0.111	1.309
	CD*A	2116	0.192	1.313
400	C*DE	1362	0.0464	2.440
	DC*A	1714	0.0429	2.502
	CD*A	1966	0.0922	2.368
500	C*AE	2088	0.0308	3.459
	C*C*A	1900	0.0279	3.555

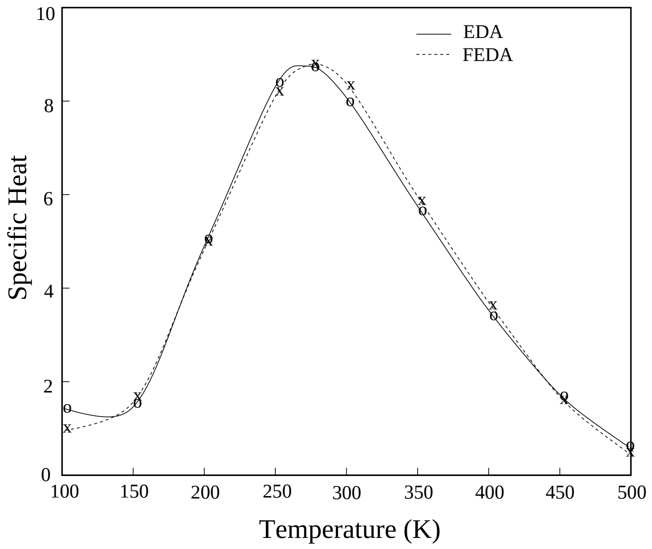

Figure 20: Plot of specific heat using EDA and FEDA free energy results for isolated met-enkephalin.

Here N refers to the number of amino acid residues in the peptide. The average energy and squared energy ($< E >_T$ and $< E^2 >_T$, respectively) were calculated at 10 temperature points using expressions of the form given in Equation (49). The bin probabilities were based on an energy width of .015625 kcal/mol. In addition, a reference free energy, G_o^{approx} (the lowest free energy), was used to normalize the probabilities at each temperature point.

The results for isolated met-enkephalin are shown in Figure 20. Both the EDA and FEDA predict a transition temperature in the 250-275 K temperature range. This is consistent with the increase in bin density and structural diversity at higher temperatures, and suggests a sharp increase in the average potential energy of the system at this temperature. It also supports the transition from the DC*B ground state (PEGM) cluster to the higher potential energy CD*A cluster in this temperature range.

Similar results for characterizing the folding transitions of enkephalins have also been obtained by multicanonical simulations [93]. This is encouraging because the two methods possess fundamental differences. In contrast to this work, the multicanonical approach does not rely on the identification of low energy local minima or the concepts of the harmonic approximation. Instead, thermodynamic quantities are developed by first generating large ensembles of structures with wide ranging energies and then employing reweighting techniques. In addition, although the multicanonical simulations included detailed atomistic level modeling, only unsolvated systems were considered.

The EDA was then applied to the RRIGS solvated form of met-enkephalin using the same protocol and conditions as detailed above. Qualitatively, the PEGM (in this case PEGM refers to potential+solvation) for solvated met-enkephalin exhibits a more extended conformation than that which is observed for the isolated form. As detailed in Table 18, the PEGM structure persists

Table 18: Dihedral angle values for PEGM and FEGM structures of solvated met-enkephalin. The temperatures are provided in the first row. The last two rows indicate the harmonic free energy (kcal/mol) and the potential energy value (kcal/mol), respectively.

Residue	DA	PEGM	100 K	200 K	300 K	400 K	500 K
Tyr_1	ϕ	-168.2	-168.2	-170.9	-168.4	-168.4	-152.5
	ψ	-30.9	-30.9	-28.5	-34.3	-34.3	153.2
	ω	178.6	178.6	177.5	-178.9	-178.9	178.5
	χ_1	-173.5	-173.5	178.8	178.7	178.7	-179.0
	χ_2	-100.9	-100.9	61.3	-100.8	-100.8	-101.2
	χ_3	19.3	19.3	-4.1	179.0	179.0	-179.9
Gly_2	ϕ	78.5	78.5	73.8	177.8	177.8	-173.9
	ψ	-86.5	-86.5	47.6	-179.9	-180.0	177.1
	ω	-177.3	-177.3	-179.2	180.0	180.0	-179.8
Gly_3	ϕ	162.4	162.4	167.6	-180.0	-180.0	179.6
	ψ	92.2	92.2	-145.2	179.9	179.9	-179.3
	ω	172.6	172.6	175.2	179.7	179.7	179.6
Phe_4	ϕ	-150.3	-150.3	-149.3	-155.3	-155.4	-155.4
	ψ	159.8	159.8	135.8	147.2	149.5	149.3
	ω	-178.1	-178.1	-176.6	-176.8	-178.3	-178.3
	χ_1	65.8	65.8	177.3	-179.5	-179.5	-179.7
	χ_2	-87.4	-87.4	-108.1	-111.7	-105.6	74.4
Met_5	ϕ	-75.0	-75.0	-85.5	-78.7	-78.7	-78.9
	ψ	113.9	113.9	-41.1	-51.1	113.4	113.5
	ω	-178.4	-178.4	179.9	179.7	-179.1	-179.1
	χ_1	-172.3	-172.3	-65.6	-67.2	-67.4	-67.4
	χ_2	176.1	176.1	-179.6	-178.8	-178.8	-178.8
	χ_3	-180.0	-180.0	-179.4	-179.9	-179.9	-179.9
	χ_4	60.0	60.0	179.5	-180.0	60.0	-60.0
G		-50.060	-41.896	-34.566	-28.604	-22.828	-17.166
E		-50.060	-50.060	-48.676	-46.030	-45.780	-44.797

as the FEGM at 100 K. However, at each subsequent temperature, the FEGM structure changes, and this change is accompanied by an increase in total energy (potential and solvation). As with isolated met-enkephalin, the difference in total energy between the PEGM and FEGM at 500 K is greater than 5 kcal/mol. This suggests that entropic effects are important in defining the predicted native structure. When considering individual structures, entropic effects tend to produce more extended FEGM conformations at higher temperatures, especially with regard to the placement of the aromatic rings. It is interesting to note that in a previous study the positioning of aromatic rings was found to be a major difference when considering the ability of solvation models to predict extended PEGM conformations for the solvated enkephalin peptides [128]. The sequence of FEGM structures is illustrated in Figure 21.

The distribution of the 72784 distinct minima for solvated met-enkephalin exhibits some impor-

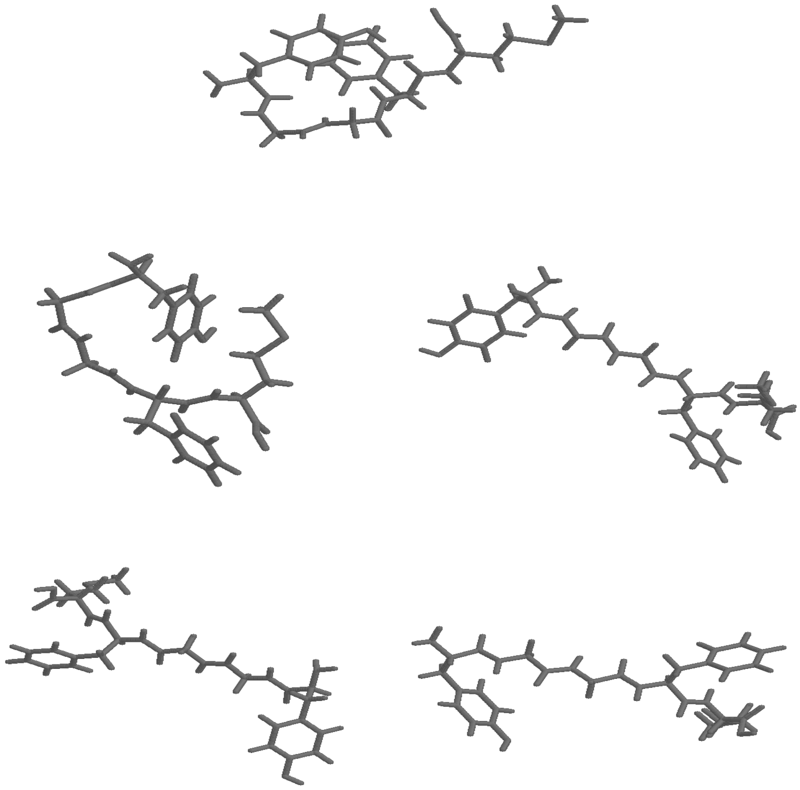

Figure 21: FEGM structures for solvated met-enkephalin. The top figure is the PEGM and the FEGM for 100 K. The structures at other temperatures (200,300,400,500) are shown left to right, top to bottom.

Table 19: Number of distinct minima in bins for solvated met-enkephalin. Each bin represents a 0.5 kcal/mol range above the previous bin. The temperatures are given in the first row.

Bin	0 K	50 K	100 K	150 K	200 K	250 K	300 K	350 K	400 K	450 K	500 K
1	10	11	16	17	21	18	19	22	21	21	13
2	14	17	35	122	236	149	98	95	97	94	79
3	34	66	299	542	896	607	378	283	223	195	166
4	117	296	668	1589	2075	1496	885	635	520	412	343
5	326	626	1907	3163	3636	2644	1730	1175	814	678	548
6	717	1582	3324	4902	5438	4256	2812	1957	1418	1047	762
7	1440	2865	5393	6733	6816	5790	4451	3061	2172	1623	1202
8	2611	4521	6906	7692	7569	6730	5390	4376	3123	2299	1705
9	3891	6337	7857	7952	7650	7221	6301	4972	4073	3132	2263
10	5567	7342	8094	7304	6858	7158	6736	5925	4699	3788	2903
11	6677	8090	7193	6612	6320	6374	6675	6232	5426	4453	3501
12	7624	7483	6618	5915	5645	6028	6295	6270	5754	5015	4161
13	7650	6920	5726	4864	4582	5279	5756	5972	5822	5328	4577
14	7047	6106	4680	3875	3645	4280	5113	5546	5689	5387	4879
15	6375	5066	3710	3086	2978	3449	4361	4973	5376	5271	5012
16	5534	4090	2848	2237	2140	2796	3437	4233	4809	5141	4964

tant differences from those results obtained for the isolated form of the peptide. This is evidenced by the information presented in Table 19 and the plot in Figure 22. In particular, the low and intermediate energy bins are much denser than the corresponding bins for isolated met-enkephalin, especially within 4 kcal/mol (8 bins) of the FEGM. In addition, some higher energy bins are actually more populated at lower temperatures. One obvious reason for these differences is the high density of conformers for the original system (at 0 K). This high density of states causes the original energy differences to be relatively small, and the entropic correction tends to induce an even stronger equalization of the free energy values. This equalization is best illustrated by the data plotted in Figure 22, which indicates that the efficiency of locating low free energy conformers is relatively high at all temperatures. In fact, the highest density of states occurs near the middle of the temperature range, rather than at high temperatures as predicted for the isolated peptide. This behavior may be due to a lack of much higher energy local minima which would probably populate these high temperature, high energy bins.

Similar conclusions can be drawn by examining the data presented in Figure 23, which provides information on the energy extrema for free energy bins at temperatures of 0, 250 and 500 K. As expected, for both 250 and 500 K, the range of energy values increases for higher free energy bins. In addition, for all bins, the minimum energy is relatively low and generally within a few kcal/mol of the PEGM. However, unlike the isolated met-enkephalin results, the maximum energy values do not become larger at higher temperatures. In fact, the curves for maximum energy at 250 and 500 K are almost identical. This indicates that relatively high energy minima may be needed in order to fill out these high temperature bins.

Table 20: Clustered relative free energies for solvated met-enkephalin. From left to right, the information provided in this table includes : temperature, Zimmerman code[i], number of individual structures in cluster, total probability ($\sum_i p_i^{approx}$) and free energy of cluster ($G_{cluster}$).

Temp (K)	Code	Number	$\sum_i p_i^{approx}$	$G_{cluster}$
	C*H*E	139	0.466	0.152
100	C*DF	286	0.224	0.297
	C*G*A	205	0.0991	0.459
	C*A*E	1112	0.0521	1.174
200	A*E*E	393	0.0468	1.217
	E*EE	149	0.0421	1.259
	E*EE	148	0.0474	1.818
300	EE*E	152	0.0445	1.856
	D*E*E	149	0.0273	2.147
	EE*E	151	0.0476	2.419
400	E*EE	145	0.0391	2.575
	EEE	159	0.0266	2.883
	EE*E	149	0.0460	3.059
500	E*EE	142	0.0327	3.397
	EEE	156	0.0299	3.488

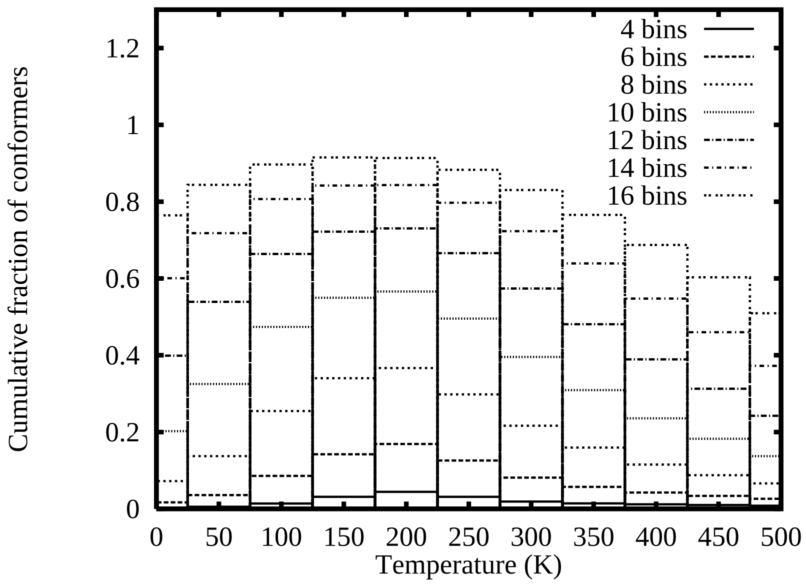

Figure 22: Plot of cumulative fraction of low energy conformers for solvated met-enkephalin, which is equal to the number of unique conformers within the first 4, 6, 8, 10, 12, 14 and 16 energy bins over the total number unique conformers, versus temperature.

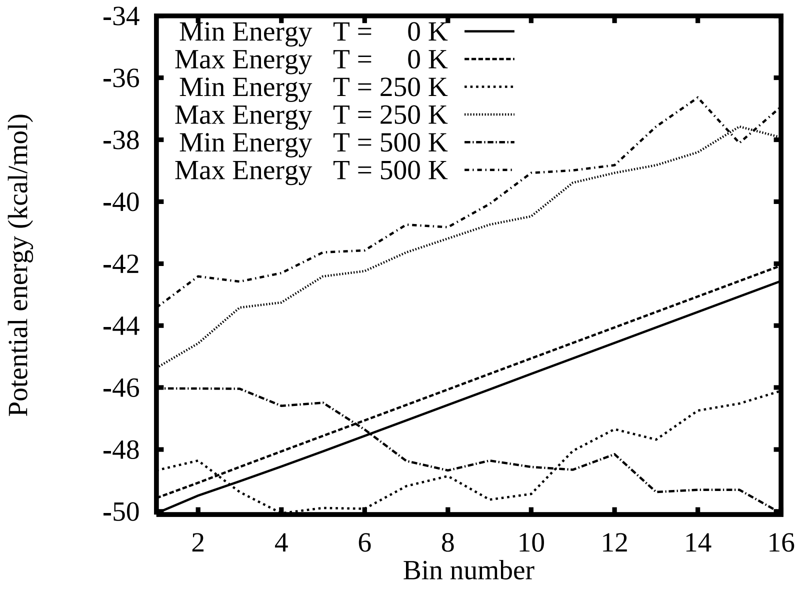

Figure 23: Energy comparison for solvated met-enkephalin. Minimum and maximum potential energies versus bin number are plotted for three temperatures : T = 0 K, 250 K, 500 K.

A clustering analysis of the low free energy conformers was also performed for solvated met-enkephalin, and the results are shown in Table 20. At 100 K, the lowest free energy cluster included the FEGM structure, which is also the PEGM structure. At higher temperatures, the correlation between the extended FEGM structures and the lowest free energy cluster was also evident. In fact, all low energy clusters at 300, 400 and 500 K possess highly extended backbone conformations, with nearly all geometries within the E and E* regions on the Zimmerman conformational map. In fact, although the number of individual structures in each cluster is not excessively large, many of these extended conformers reside in the lowest free energy bins.

A specific heat profile was also derived for solvated met-enkephalin in order to understand how the dominance of these extended cluster geometries affect the folding transition. These results are shown in Figure 24. As with isolated met-enkephalin, a folding transition is indicated by the peak in the specific heat, which, in this example, occurs between 275 and 300 K. This represents a significant change in average energy, which accompanies the collapse from an ensemble of extended conformations (EE*E and E*EE clusters) to the more compact ground state cluster. For the solvated met-enkephalin example, this transition is clearly illustrated by the cluster analysis and the structure plots given in Figure 21.

B Structure Refinement with Sparse Restraints

To effectively determine protein function it is important to predict the three dimensional structure of the macromolecule. Over the last several decades a number of experimental and theoretical approaches have been developed and refined in order to achieve this goal. Experimentally, there

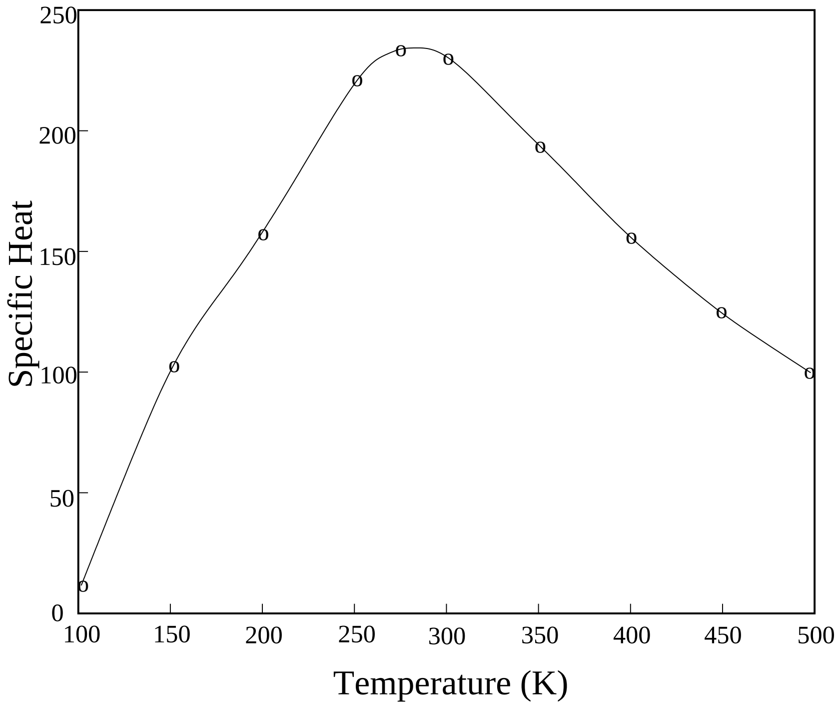

Figure 24: Plot of specific heat using free energy results for solvated met-enkephalin.

now exist two basic techniques used to perform protein structure refinement. The first, X-ray crystallography, relies on the ability to crystallize the protein so that diffraction patterns can be used for sufficient resolution. These requirements have limited the applicability of this technique. A more powerful method, NMR (nuclear magnetic resonance) spectroscopy, is based on solution measurements of the system. Several key developments, including multidimensional NMR experiments, have resulted in the ability to determine solution structures for proteins consisting of over 200 residues.

This section focuses on the development of a novel approach for protein structure prediction via experimental NMR restraints. Traditionally, the protein folding global optimization problem involves a progression of unconstrained minimizations. However, the introduction of experimentally derived or artificial restraints can be used to recast the fundamental protein folding problem as a constrained global optimization problem. 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.

This constrained approach is applied to the NMR structure prediction problem, although a variety of restraint information could be used. The proposed constrained formulation differs from traditional NMR approaches in several fundamental ways. First, the energy model is represented by a detailed full atom force field, rather than simplified nonbonded potential terms. This should make the approach especially effective when the number of NMR restraints per residue decreases; that is, the accuracy of the energy model becomes more significant. In addition, traditional solution approaches apply target function distance geometry or simulated annealing to unconstrained problem formulations in which restraints are represented by penalty function approximations. The solution of the constrained formulation requires the use of constrained local optimization solvers and an overall global optimization framework for nonlinearly constrained problems. This is accomplished through the application of the ideas of the αBB deterministic global optimization approach [3, 1, 4, 2, 9]. αBB based global optimization techniques have also been applied to NMR type structure prediction problems [132, 220].

Because the location of the global minimum relies on effectively solving constrained local optimization problems, convergence to the global minimum can be enhanced by consistently identifying low energy solutions. These observations illustrate the need for reliably locating low energy feasible points for initializing the constrained local optimization routine. Along these lines, a combined torsion angle dynamics (TAD) and simulated annealing scheme has been implemented within the context of the global optimization framework. Torsion angle dynamics (TAD) has recently been shown to be more effective than Cartesian coordinate dynamics [91, 198]. In this case, the degrees of freedom are rotations around single bonds, which reduces the number of variables by approximately tenfold because bond lengths, bond angles, chirality and planarities are kept fixed at optimal values during the calculation.

1 Energy modeling

Basic data obtained from NMR studies consist of distance and torsion angle restraints. Once resonances have been assigned, nuclear Overhauser effect (NOE) contacts are selected and their intensities are used to calculate interproton distances. Information on torsion angles are based on the measurement of coupling constants and analysis of proton chemical shifts. Together, this information is used to formulate a nonlinear optimization problem, the solution of which should provide the correct protein structure. Typically, a hybrid energy function of the following form is employed:

$$E = E_{\text{forcefield}} + W_{\text{nmr}} E_{\text{nmr}}. \tag{57}$$

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

The second term of Equation (57) can be rewritten as :

$$E_{\text{nmr}} = E_{\text{distance}} + E_{\text{dihedral}}. \tag{58}$$

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. The expressions also include a summation over both upper and lower distance violations; for example, $E_{\text{distance}} = E_{\text{distance}}^{\text{upper}} + E_{\text{distance}}^{\text{lower}}$. When considering upper distance restraints this becomes:

$$E_{\text{distance}}^{\text{upper}} = \sum_j \begin{cases} A_j(d_j - d_j^{\text{upper}})^2 & \text{if } d_j > d_j^{\text{upper}}, \\ 0 & \text{otherwise.} \end{cases} \tag{59}$$

The squared violation energy is considered only when the calculated distance d_j exceeds the upper reference distance d_j^{upper}. This squared violation can then multiplied by a weighting factor A_j. A similar contribution is calculated for those distances that violate a lower reference distance, d_j^{lower}:

$$E_{\text{distance}}^{\text{lower}} = \sum_j \begin{cases} A_j(d_j - d_j^{\text{lower}})^2 & \text{if } d_j < d_j^{\text{lower}}, \\ 0 & \text{otherwise.} \end{cases} \tag{60}$$

For dihedral angle restraints the functional form is similar to that of Equations (59) and (60). As before, the total violation, E_{dihedral}, is a sum over upper and lower violations (i.e., $E_{\text{dihedral}} = E_{\text{dihedral}}^{\text{upper}} + E_{\text{dihedral}}^{\text{lower}}$). A dihedral angle ω_j can be restrained by employing a quadratic square well potential using upper (ω_j^{upper}) and lower (ω_j^{lower}) 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. Furthermore, by centering the full periodic region on the region defined by the allowable bounds, all transformed values will lie in the domain defined by $[\omega_j^{\text{lower}} - \Delta HW_{\omega_j}, \omega_j^{\text{upper}} + \Delta HW_{\omega_j}]$, where ΔHW_{ω_j} is equal to half the excluded range of dihedral angle values (i.e., $\Delta HW_{\omega_j} = \pi - (\omega_j^{\text{upper}} - \omega_j^{\text{lower}})/2$). This results in the following set of equations:

$$E_{\text{dihedral}}^{\text{upper}} = \sum_j \begin{cases} A_j \left(1 - 2\left[\frac{\omega_j - \omega_j^{\text{upper}}}{2\pi - (\omega_j^{\text{upper}} - \omega_j^{\text{lower}})}\right]^2\right)(\omega_j - \omega_j^{\text{upper}})^2 & \text{if } \omega_j > \omega_j^{\text{upper}}, \\ 0 & \text{otherwise,} \end{cases} \tag{61}$$

$$E_{\text{dihedral}}^{\text{lower}} = \sum_j \begin{cases} A_j \left(1 - 2\left[\frac{\omega_j - \omega_j^{\text{lower}}}{2\pi - (\omega_j^{\text{upper}} - \omega_j^{\text{lower}})}\right]^2\right)(\omega_j - \omega_j^{\text{lower}})^2 & \text{if } \omega_j < \omega_j^{\text{lower}}, \\ 0 & \text{otherwise.} \end{cases} \tag{62}$$

The force field energy term, $E_{\text{forcefield}}$ of Equation (57), models the nonbonded interactions of the protein, which can consist of simple or more detailed energy functions. In practice, when considering NMR restraints, force field terms are often simplified to include only geometric energy terms such as quartic Van der Waals repulsions. Such functions neglect rigorous modeling of energetic terms in order to ensure that experimental distance violations are minimized. In fact, a basic representation for this target function would be :

$$E_{\text{S}} = E_{\text{distance}} + E_{\text{dihedral}}. \tag{63}$$

Here the E_{distance} function includes additional distance restraints to avoid van der Waals contacts. Notice that when all restraints are satisfied, the objective function is driven to zero.

A detailed modeling approach is proposed by using the ECEPP/3 force field [180]. When considering an unconstrained minimization, this approach provides the following objective function :

$$E_{\text{D}} = E_{\text{distance}} + E_{\text{dihedral}} + E_{\text{ECEPP/3}}. \tag{64}$$

In contrast to Equation (63), 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.

2 Global optimization

The determination of a three dimensional protein structure defines an optimization problem in which the objective function may correspond to one of the target functions outlined in the previous section. For the simple case, the formulation becomes :

$$\min_{\phi} \qquad E_{\mathrm{S}}(\phi) = E_{\mathrm{distance}} + E_{\mathrm{dihedral}}. \tag{65}$$

A standard procedure for addressing this global optimization problem consists of a combination of simulated annealing and molecular or torsional angle dynamics [32]. Generally, multiple initial conformers are generated and optimized to provide a set of acceptable structures. Typically, a set containing on the order of 100 acceptable conformers may be identified, from which a subset of similar structures (approximately 20) are used to characterize the system. The simulated annealing protocol is incorporated in an attempt to reduce trapping in local minimum energy wells.

However, the minimization of complex target functions necessitates the use of rigorous global optimization techniques. For the detailed target function, given by Equation (64), the unconstrained formulation is similar to formulation (65). Through the use of the constrained optimization approach, the dihedral angle bounds are implicitly included as box constraints. Furthermore, distance restraints are treated explicitly. This reformulation removes both E_{dihedral} and E_{distance} from the target function, leaving only $E_{\mathrm{forcefield}}$:

$$\begin{aligned} \min_{\phi} \qquad & E_{\mathrm{ECEPP/3}}, \\ \text{subject to} \qquad & E_l^{\mathrm{distance}}(\phi) \;\leq\; E_l^{\mathrm{ref}} \quad l = 1, \ldots, N_{\mathrm{CON}}, \\ & \phi_i^L \;\leq \phi_i \leq\; \phi_i^U, \;\; i = 1, \ldots, N_{\phi}. \end{aligned} \tag{66}$$

Here $i = 1, \ldots, N_{\phi}$ 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 derived from NMR data are also suitable.

3 Torsion angle dynamics

Standard unconstrained molecular dynamics simulations have been used extensively to model the folding and unfolding of protein systems [62, 52, 37]. In addition, several methods for NMR structure calculation have been based on molecular dynamics in Cartesian space [32]. 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 [91, 198].

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 [107].

The algorithm solves for the torsional accelerations, $\ddot{\theta}$:

$$M(\theta)\ddot{\theta} \; + \; \mathbf{C}(\theta,\dot{\theta}) \; = \; \mathbf{0}. \tag{67}$$

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. θ, $\dot{\theta}$ and $\ddot{\theta}$ represent the torsional position, velocities and accelerations, respectively. The ability to calculate the accelerations recursively relies on the chainlike structure of the protein, in 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 θ_k.

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

Step 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, $\phi_{\mathbf{k}}$, between rigid bodies k and $k-1$ must first be defined :

$$\phi_{\mathbf{k}} = \begin{bmatrix} I_3 & \tilde{l}(\mathbf{r_k} - \mathbf{r_{k-1}}) \\ 0_3 & I_3 \end{bmatrix}. \tag{68}$$

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 :

$$\mathbf{V_k} = \phi_\mathbf{k}^\mathbf{T}\mathbf{V_{k-1}} + \mathbf{H_k^T}\dot{\theta}_\mathbf{k}. \tag{69}$$

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

$$\mathbf{V_k} \equiv \begin{pmatrix} \omega_\mathbf{k} \\ \mathbf{v_k} \end{pmatrix}. \tag{70}$$

$\mathbf{H_k}$ 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 $k-1$:

$$\mathbf{H_k} \equiv \begin{pmatrix} \mathbf{e_k} \\ 0 \end{pmatrix}. \tag{71}$$

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

$$\mathbf{a_k} = \begin{pmatrix} 0 \\ \tilde{\omega}_{k-1}[\mathbf{v_k} - \mathbf{v_{k-1}}] \end{pmatrix} + \begin{pmatrix} \tilde{\omega_\mathbf{k}} & 0 \\ 0 & \tilde{\omega_\mathbf{k}} \end{pmatrix}\mathbf{H_k^T}\dot{\theta}_\mathbf{k} \tag{72}$$

$$\mathbf{b_k} = \begin{pmatrix} \tilde{\omega}_k\mathbf{J_k}\tilde{\omega}_\mathbf{k} \\ m_k\tilde{\omega}_k\tilde{\omega}_k\mathbf{Y_k} \end{pmatrix}. \tag{73}$$

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:

$$\mathbf{L_k} = \begin{pmatrix} \mathbf{J_k} & m_k\tilde{\mathbf{Y}}_\mathbf{k} \\ -m_k\tilde{\mathbf{Y}}_\mathbf{k} & m_kI_3 \end{pmatrix}. \tag{74}$$

Step 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 :

$$D_k = \mathbf{H_k}\mathbf{P_k}\mathbf{H_k^T} \tag{75}$$

$$\mathbf{G_k} = \mathbf{P_k}\mathbf{H_k^T}\mathbf{D_k^{-1}} \tag{76}$$

$$\epsilon_k = -\mathbf{H_k}(\mathbf{z_k} + \mathbf{P_k}\mathbf{a_k}) - \nabla\mathbf{E_k}. \tag{77}$$

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 :

$$\mathbf{P_{k-1}} \leftarrow \mathbf{P_{k-1}} + \phi_\mathbf{k}(\mathbf{P_k} - \mathbf{G_k}\mathbf{H_k^T}\mathbf{P_k})\phi_\mathbf{k}^\mathbf{T} \tag{78}$$

$$\mathbf{z_{k-1}} \leftarrow \mathbf{z_{k-1}} + \phi_\mathbf{k}(\mathbf{z_k} + \mathbf{P_k}\mathbf{a_k} + \mathbf{G_k}\epsilon_\mathbf{k}). \tag{79}$$

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

$$\alpha_{\mathbf{k}} = \phi_{\mathbf{k}}^{\mathbf{T}} \alpha_{\mathbf{k-1}} \tag{80}$$

$$\ddot{\theta}_k = \epsilon_k D_k^{-1} - \mathbf{G}_{\mathbf{k}} \alpha_{\mathbf{k}}. \tag{81}$$

The following recursion relation is used to update the values of $\alpha_{\mathbf{k}}$

$$\alpha_{\mathbf{k}} \leftarrow \alpha_{\mathbf{k}} + \mathbf{H}_{\mathbf{k}} \ddot{\theta}_{\mathbf{k}} + \mathbf{a}_{\mathbf{k}}. \tag{82}$$

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 [23]. This coupling provides a means for forcing or damping the torsional velocities using the following scaling factor at time t:

$$f_T = \sqrt{1 - \frac{1}{\beta} + \frac{T_o}{\beta T(t)}}. \tag{83}$$

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 :

$$T(t) = \frac{2E_{\text{kinetic}}(t)}{N_\phi k_{\text{B}}}. \tag{84}$$

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

$$\dot{\theta}(\mathbf{t}) \leftarrow \mathbf{f_T} \dot{\theta}(\mathbf{t}). \tag{85}$$

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

4 Algorithmic steps

The algorithmic steps for the constrained αBB approach can be generalized to any force field model or routine for solving constrained optimization problems. Here, the αBB approach is interfaced with PACK [205] and NPSOL [79]. PACK is used to transform to and from Cartesian and internal coordinate systems, and 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 can be broken down into two main 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 based on their overall effect on the structure of the molecule. In this work (and in general) the ϕ and ψ dihedral angles provide the largest structural variability, and are chosen to constitute the global variable set.

(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 [-π,π]. Since a constrained local optimization solver is used these box constraints are strictly enforced.

(3) Identify the set of NOE derived distance restraints to be used in the constraints. In general, this set can include all intra- and inter-residue restraints. In this work, only backbone sequential and medium/long range information was used in developing the constraints, because intra-residue restraints are less likely to affect the overall fold. 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. In this work, information based on X-PLOR [32] results was used to define the E^{ref} parameter (see below).

(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 of the algorithm involves an iterative approach, which depends on the refinement of the original domain by 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) 100 random points are generated and used for evaluation of the objective function and constraints. The point with the minimum objective function value and feasible constraints is used as a starting point for local minimization using NPSOL. If a feasible starting point is not found, local minimization is not performed.

(B) All feasible solutions are stored.

(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 $L_{\text{forcefield}}$ 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.

To enhance the search for low energy feasible points, the basic procedure described in Step 3.A is modified to include TAD. The following protocol is used :

(1) Set counter, $c = 1$. Perform TAD (1000 high temperature steps followed by 3000 annealing steps) using E_{S} as the 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.

(A) If the $E_l^{\text{distance}} < E_l^{\text{ref}} \;\; \forall l = 1, \ldots, N_{\text{CON}}$, go to step 2. Else go to step 1.B.

(B) Increment counter, $c = c + 1$. If $c < 5$, reduce weight of sterically based distance restraints, perform new TAD and go to step 1.A. Else go to step 2.

(2) Set counter, $c = 1$. Perform local minimization using NPSOL with dihedral angle box constraints to implicitly enforce bounds. The objective function is a weighted combination of forcefield energy and distance restraint terms :

$$E = E_{\text{ECEPP/3}} + \sum_l W_l E_l^{\text{distance}}. \tag{86}$$

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

$$W_l = \sqrt{1 + \frac{E_l^{\text{distance}}}{E_l^{\text{ref}}}}. \tag{87}$$

(A) If $E_l^{\text{distance}} < E_l^{\text{ref}} \;\; \forall l = 1, \ldots, N_{\text{CON}}$, go to step 3. Else go to step 2.B.

(B) 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 2.A. Else go to step 3.

(3) Solve the constrained minimization problem using NPSOL.

5 Computational study

The global optimization algorithm was tested on Compstatin, a synthetic 13-residue (ICVVQD-WGHHRCT) cyclic peptide (disulfide bridge between the Cys^2 and Cys^{12} residues) that binds to C3 (third component of complement) and inhibits complement activation [203]. Two-dimensional NMR techniques [177] yield a total of 30 backbone sequential (including H^β – backbone), 23 medium and long range (including disulfide) and 82 intra-residue NOE restraints. In addition, 7 ϕ angle and 2 χ_1 angle dihedral restraints are available. In previous work [177], traditional distance geometry–simulated annealing protocol was utilized to minimize the associated target function in the Cartesian coordinate space using the program X-PLOR [32]. NOE distance and dihedral angle restraints were modeled using a quadratic square well potential, while van der Waals overlaps were prevented through the use of a simple quartic potential function.

The NMR refinement protocols resulted in a family of 21 structures with similar geometries for the Gln^5–Gly^8 region. A representative structure was obtained by averaging the coordinates of the individually refined structures and then subjecting this structure to further refinement to release geometric strain produced by the averaging process. The formation of a type I β-turn was identified as a common characteristic for these structures.

The consistency of the ensemble of Compstatin solution structures was determined by evaluating distance restraints for each of the original 21 structures (accession number 1a1p at the Brookhaven Protein Data Bank, http://www.pdb.bnl.gov), as well as the average Compstatin conformation. In considering distance restraints, only backbone sequential and medium/long range NOEs were considered. That is, the 82 intra-residue restraints were neglected since they are less likely to effect the overall fold of the Compstatin peptide. This results in a total of 52 restraints, with an additional restraint on the distance between the sulfur atoms forming the disulfide bridge. In order to quantify these results the violation energy, E_{VIO}, which can be calculated by summing Equations (59) and (60), was calculated for each of the original PDB structures. In these calculations, the value of the weighting factor (A_j) was assumed to be constant and set equal to 50 kcal/mol/$\mathring{A}^2$.

The results of the analysis, shown graphically in Figure 25, indicate that the average structure ($\overline{Compstatin}$) possesses the third largest violation energy, whereas the smallest value is provided by structure 8 ($< Compstatin >_8$). These quantities provide a range of comparison for violation energies and were used to set the constraint parameter, E^{ref}, to 200 kcal/mol. This value is chosen so that the sum of the violation energies will necessarily result in an improvement over the violation energy for the average Compstatin structure.

To measure the performance of the proposed global optimization approach, the ensemble and average Compstatin structures ($< Compstatin >$ and $\overline{Compstatin}$) were then used as starting points for local minimization. Since these calculations are performed in the torsion angle space, which requires fixing bond lengths and bond angles to equilibrium values, the corresponding Compstatin PDB structures could only be used to derive torsion angle values. These dihedral angles were then used as input to directly evaluate the corresponding force field energy. Differences in bond lengths and bond angles propagate through the generation of the corresponding ECEPP/3 structure, which produces an inherent RMSD between the PDB structure and the ECEPP/3 generated

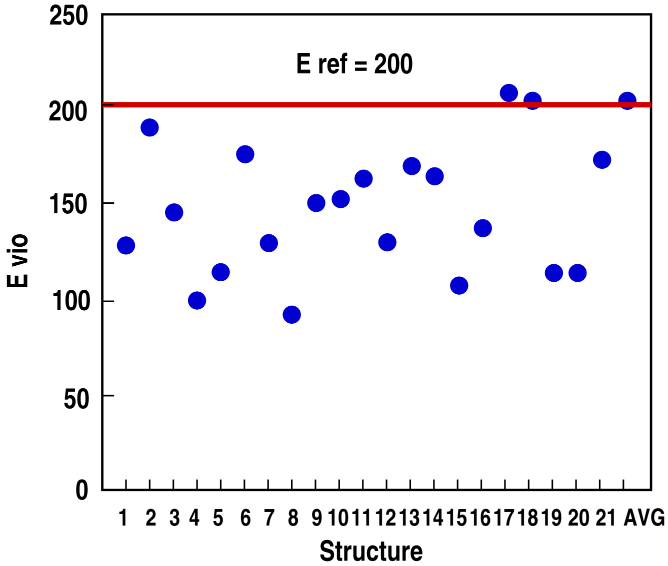

Figure 25: Violation energy, E_{VIO}, for original Compstatin PDB structures.

structure. For example, when using the set of dihedral angles calculated from the $\overline{Compstatin}$ PDB, the ECEPP/3 structure possesses a 0.581 Å all atom RMSD (all heavy atoms in backbone and side chains) with respect to the original $\overline{Compstatin}$ structure, with a corresponding ECEPP/3 energy of 519.2 kcal/mol. In addition, due to the differences in bond lengths and angles, the total distance violation for the ECEPP/3 structure ($\overline{Compstatin}_{ECEPP}$) increases from 6.9 to 8.7 Å, which results in a subsequent increase in violation energy to 315 kcal/mol. The superposition of the original and ECEPP/3 $\overline{Compstatin}$ conformations is shown in Figure 26.

Due to the relatively large distance violations and energies obtained after transformation of PDB to PACK (ECEPP/3) structures, the 22 structures were then subjected to local minimization. The problem formulation for local minimization uses the set of 53 restraints for the constraint function, a constant 50 kcal/mol/Å weighting factor (A_j) and a constraint parameter, E^{ref}, equal to 200 kcal/mol. In all cases, the corresponding violation energy reached the upper bound value of 200 kcal/mol. The corresponding total distance violations increased, with an average value of 6.766 Å. The smallest distance violation (5.873 Å) was reported for structure number 10 ($< Compstatin >_{10}^{Local}$), whereas the corresponding energy for this structure (-41.685 kcal/mol) was only slightly above the average energy of -47.75 kcal/mol. The lowest energy structures (-71.613 for $< Compstatin >_{2}^{Local}$, -68.704 kcal/mol for $< Compstatin >_{21}^{Local}$, -67.653 kcal/mol for $< Compstatin >_{9}^{Local}$) provided above average values for total distance violation (6.963 Å, 6.832 Å, 7.120 Å, respectively). In addition, the conformation obtained from the average Compstatin structure ($\overline{Compstatin}$) exhibited near average values for energy (-52.283 kcal/mol) and total distance violations (6.392 Å). The range of ECEPP/3 energies after local minimization are shown in Figure 27.

The structural characteristics of these locally minimized structures were quantified using RMSD

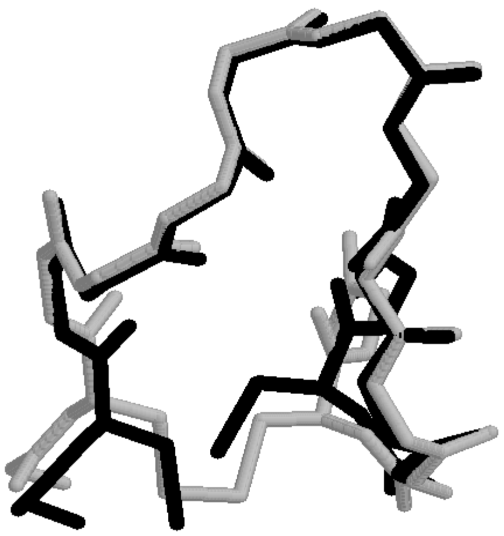

Figure 26: Superposition of $\overline{Compstatin}_{Orig}$ structure (in light grey) and corresponding ECEPP/3 structure (in black) using calculated dihedral angles ($\overline{Compstatin}_{ECEPP}$).

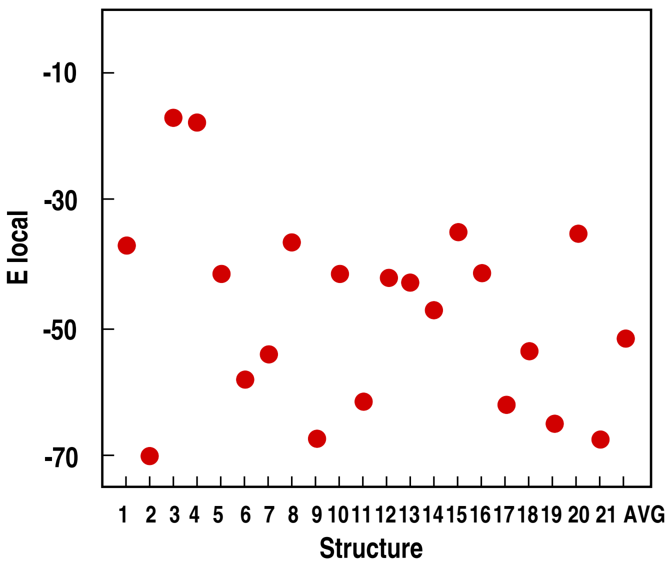

Figure 27: Locally minimized energy, $E_{\mathrm{ECEPP/3}}$, for Compstatin structures.

(root mean squared deviation) calculations. For the original PDB structures, comparison with the average Compstatin structure provided RMSD values between $1-2\AA$ for only backbone atoms. As expected, these structures possess common structural features. However, when comparing original PDB structures and their locally minimized counterparts, most RMSD values are larger than 2 $\AA$, indicating that significant conformational change occur during local minimization. This is due to both the reduced set of NOE restraints in the constraint function and the role of the detailed energy force field. In contrast, the RMSD values for the β-turn region remain consistently low when comparing the original PDB structures to their locally minimized counterparts. These results indicate that the β-turn is a conserved structural feature, even with the addition of the detailed energy model.

The constrained global optimization approach was first applied to Compstatin structure prediction without the use of TAD. A subset of 26 (all ϕ and ψ) torsion angle, from a total of 73, were treated globally, while the remaining were allowed to vary locally. As was the case for local minimization, the same set of restraints were used to formulate the nonlinear constraint, with a constant 50 kcal/mol/$\AA$ weighting factor and a constraint parameter equal to 200 kcal/mol. The lowest energy structure satisfying the constraint functions provided an ECEPP/3 energy of -85.71 kcal/mol, an energy value more than 15 kcal/mol lower than those values provided by local minimization. The global minimization required approximately 40 CPU hours on a HP C160. The total distance violation equaled 6.690 $\AA$ which is near the average distance violation for the local minimum structures.

RMSD calculations were performed to again quantify the structural differences between the global minimum energy structure and the other Compstatin structures. RMSD values between the full backbone and the β-turn segments of the 22 locally minimized PDB structures and the global minimum energy structure are plotted in Figures 28 and 29, respectively. When comparing full backbone RMSD values, the $< Compstatin >_9^{Local}$, $< Compstatin >_{21}^{Local}$, $< Compstatin >_{19}^{Local}$ and $< Compstatin >_{17}^{Local}$ provide the best agreement with the global minimum energy structure. These structures also correspond to four of the lowest energy local minimum, indicating that some of the lowest energy conformers exhibit similar backbone structural characteristics. In contrast, the lowest energy local minimum, $< Compstatin >_2^{Local}$, is less similar to the global minimum energy structure. For the β-turn segment, the correlation between low RMSD values and low energy local minima does not exist. This observation, coupled with the relatively low RMSD values between all structures, indicates that the β-turn structure is a characteristic for all conformers, including the global minimum energy structure. Plots for superpositioning (backbone atoms) of the average local minimum energy structure $\overline{Compstatin}^{Local}$ and the global minimum energy structure are given in Figure 30.

6 Comparison with TAD: DYANA

A comparison to an independent method for solving distance restraint problems was also made in order to gauge the performance of the proposed αBB constrained formulation. Specifically, a torsional angle dynamics (rather than a Cartesian coordinate dynamics such as X-PLOR) package

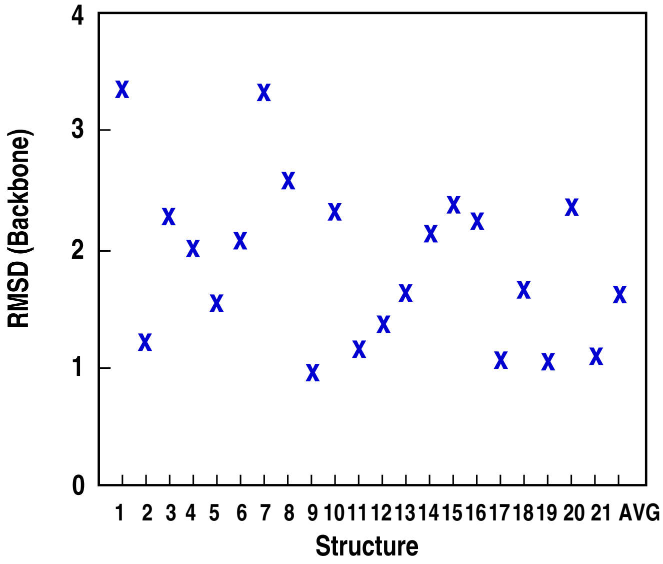

Figure 28: RMSD values for backbone when comparing global minimum energy structure to locally minimized PDB structures.

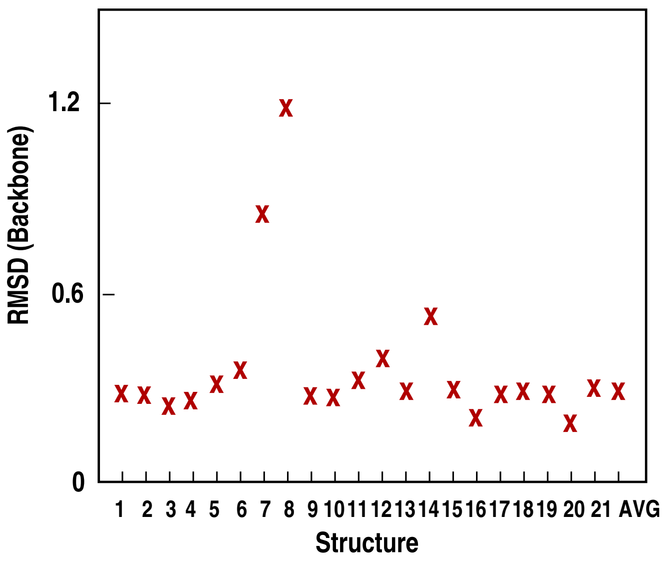

Figure 29: RMSD values for Gln^5-Gly^8 backbone when comparing global minimum energy structure to locally minimized PDB structures.

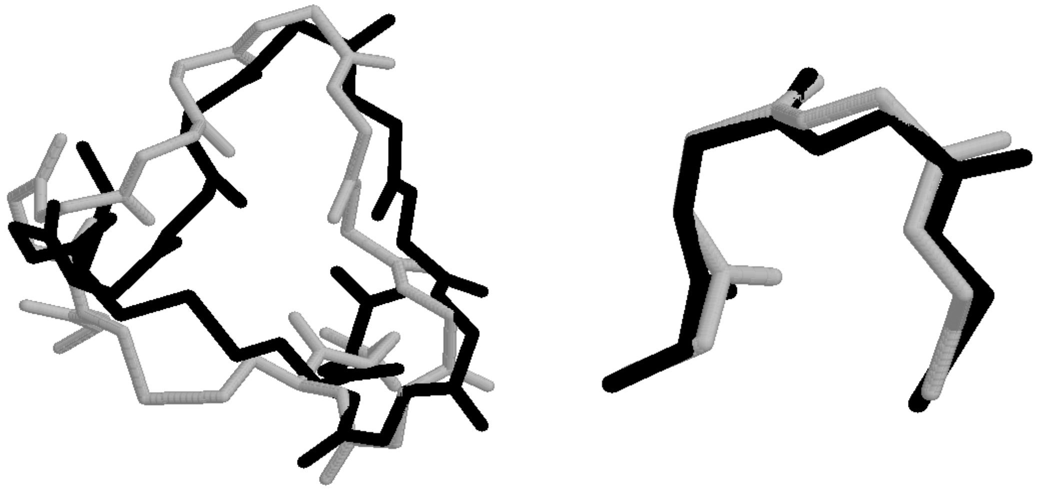

Figure 30: Superposition of global minimum (in black) and $\overline{Compstatin}^{Local}$ (in light grey) structures. The left panel shows the full (backbone atom) structure, while the right panel compares only the β-turn region.

was used [91]. The coupled simulated annealing / TAD protocol from DYANA was applied to a starting sample of 1000 randomly generated structures. The same dihedral angle constraints and 53 medium and long range distance constraints were considered; that is, no heuristic methods for reducing the variable space were employed. In the case of unspecified symmetric hydrogens, a pseudoatom approach, in which the restraint is based on a pseudoatom central to the symmetric hydrogen atoms, was used. A subset consisting of the 20 conformers exhibiting the best target values were then used as starting points for a second set of runs. Finally, a set of 5 conformations (with the smallest violations) were used for further analysis. Since each method (DYANA vs. ECEPP/3) employed different structural definitions, based on fixed bond lengths and bond angles, a direct comparison was not sufficient. Instead, the DYANA generated structures were used as starting points for local minimizations using the local constrained formulation. In all cases, the violations reached the upper bound of 200 kcal/mol for E^{ref}. The corresponding violation values, including final local minimum energy values ($E_{ECEPP/3}$) are given in Table 21.

The results given in Table 21 indicate that although the DYANA conformers satisfy the corresponding constraint, their energy values are significantly higher than that of the global minimum energy structure (more than 70 kcal/mol). This can be anticipated because the goal of the DYANA algorithm is to minimize distance restraint violations via penalty term optimization, while neglecting any detailed force field terms. In fact, an analysis of the structural characteristics indicate that the type I β-turn does not appear along the Gln^5-Gly^8 backbone in these structures. This is verified by the data in Table 22, which gives the ϕ and ψ dihedral angle values for the central β-turn residues. The problem is evidenced by the Asp^6 residue, which has ϕ-ψ values in a forbidden region

Table 21: Local minimization results for the best DYANA (TAD) generated conformations. D_{VIO} refers to the total distance violation, E_{VIO} is the corresponding violation, and energy and $E_{ECEPP/3}$ is the force field energy at the local minima.

Local minimum	D_{VIO} (Å)	E_{VIO} (kcal/mol)	$E_{ECEPP/3}$ (kcal/mol)
$Compstatin_1^{DYANA}$	6.234	200.0	-11.945
$Compstatin_2^{DYANA}$	6.538	200.0	6.782
$Compstatin_3^{DYANA}$	6.163	200.0	-10.208
$Compstatin_4^{DYANA}$	5.476	200.0	-14.516
$Compstatin_5^{DYANA}$	6.927	200.0	5.006

Table 22: ϕ and ψ values for central residues (Asp^6 and Trp^7) for anticipated β-turn region. The subscripts refer to the second and third residues in the Gln^5-Gly^8 sequence.

Local minimum	ϕ_2 (^{o})	ψ_2 (^{o})	ϕ_3 (^{o})	ψ_3 (^{o})
$Compstatin_1^{DYANA}$	166.9	-66.07	-80.00	-40.40
$Compstatin_2^{DYANA}$	165.9	-65.55	-81.02	-33.99
$Compstatin_3^{DYANA}$	180.0	-60.94	-81.76	-42.43
$Compstatin_4^{DYANA}$	168.8	-50.32	-80.00	-42.22
$Compstatin_5^{DYANA}$	165.4	-72.75	-97.79	-39.86

of the Ramachandran plot. It appears that this may be related to clustering of the side chains in the DYANA predicted structures.

In order to further examine this deviation from the previous results (which define a type I β-turn) the DYANA protocol was also tested on the full set of restraints, including intra-residue distances. The 5 DYANA predicted structures exhibiting the lowest target function values were then subjected to local minimization using the constrained formulation. As before, only the 53 medium and long range distance restraints were included during the local minimizations. As the results in Table 23 show, the average energy has decreased for this set of conformers. However, the structural analysis of the Gln^5-Gly^8 region, given in Table 24 still indicates that a type I β-turn is not preferred.

An additional comparison between the structural characteristics of these (DYANA) local minima and the global minimum was also performed using RMSD calculations, as given in Tables 25 and 26. These values are consistently larger than those between the average ($\overline{Compstatin}^{Local}$) and local minimum solutions structures ($< Compstatin >_i^{Local}$), and global minimum energy structure. The RMSD values not only indicate that there is significant structural difference over the entire structure (Table 25), but also that the β-turn region (Table 26) is not a structural characteristic of the DYANA local minima. This is evidenced by the superpositioning of the lowest energy DYANA structure and the global minimum energy structure, given in Figure 31.

Table 23: Local minimization results for the best DYANA (TAD) generated conformations using all restraints. D_{VIO} refers to the total distance violation, E_{VIO} is the corresponding violation, and energy and $E_{ECEPP/3}$ is the force field energy at the local minima.

Local minimum	D_{VIO} (Å)	E_{VIO} (kcal/mol)	$E_{ECEPP/3}$ (kcal/mol)
$Compstatin_{1c}^{DYANA}$	6.222	200.0	24.714
$Compstatin_{2c}^{DYANA}$	5.643	200.0	-31.216
$Compstatin_{3c}^{DYANA}$	6.527	200.0	-17.569
$Compstatin_{4c}^{DYANA}$	7.135	200.0	-27.110
$Compstatin_{5c}^{DYANA}$	5.926	200.0	-14.656

Table 24: ϕ and ψ values for central residues (Asp^6 and Trp^7) for anticipated β-turn region. The subscripts refer to the second and third residues in the Gln^5-Gly^8 sequence.

Local minimum	ϕ_2 (^{o})	ψ_2 (^{o})	ϕ_3 (^{o})	ψ_3 (^{o})
$Compstatin_{1c}^{DYANA}$	-180.0	-58.61	-80.00	-47.72
$Compstatin_{2c}^{DYANA}$	177.5	-63.77	-82.74	-33.53
$Compstatin_{3c}^{DYANA}$	180.0	-63.98	-82.18	-23.32
$Compstatin_{4c}^{DYANA}$	163.0	-58.56	-109.2	-4.53
$Compstatin_{5c}^{DYANA}$	-180.0	-70.46	-92.40	-41.22

Table 25: RMSD values for full Compstatin structures. Column 2 reports RMSD using all heavy atoms, while column 3 accounts for only backbone atoms (N,C^α,C'). Both columns compare the DYANA local minimum structures ($Compstatin_i^{DYANA}$) to the global minimum Compstatin PDB structure ($Compstatin^{Global}$).

Local minimum	Heavy Atoms	Backbone Atoms
$Compstatin_{1c}^{DYANA}$	4.117	2.812
$Compstatin_{2c}^{DYANA}$	4.866	3.893
$Compstatin_{3c}^{DYANA}$	5.243	3.943
$Compstatin_{4c}^{DYANA}$	4.892	2.654
$Compstatin_{5c}^{DYANA}$	4.506	3.180

Table 26: RMSD values for the β-turn regions (Residues 5 through 8). Column 2 reports RMSD using all heavy atoms, while column 3 accounts for only backbone atoms (N,C^{α},C'). Both columns compare the DYANA local minimum structures ($Compstatin_i^{DYANA}$) to the global minimum Compstatin PDB structure ($Compstatin^{Global}$).

Local minimum	Heavy Atoms	Backbone Atoms
$Compstatin_{1c}^{DYANA}$	1.163	0.625
$Compstatin_{2c}^{DYANA}$	1.473	0.732
$Compstatin_{3c}^{DYANA}$	1.607	0.721
$Compstatin_{4c}^{DYANA}$	1.327	0.721
$Compstatin_{5c}^{DYANA}$	1.277	0.781

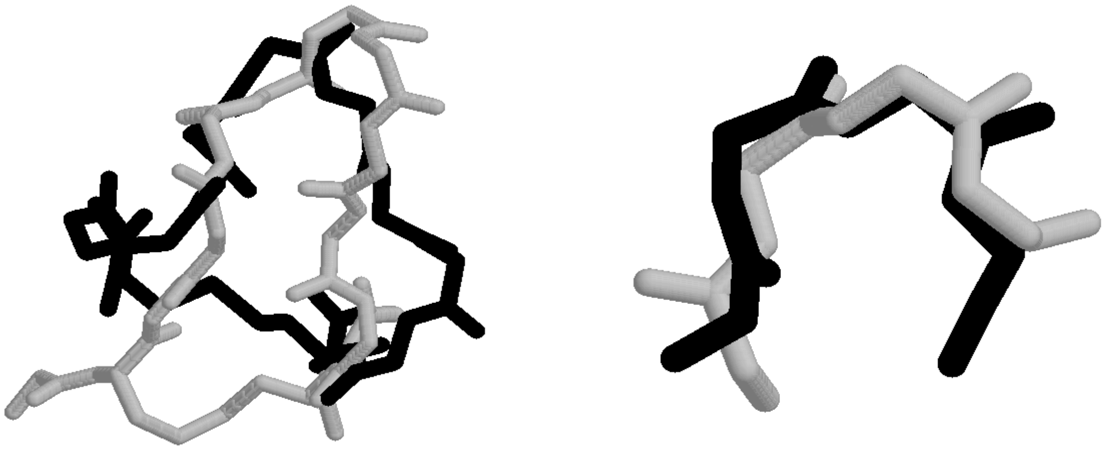

Figure 31: Superposition of global minimum (in black) and $Compstatin_{1c}^{DYANA}$ (in grey) structures. The left panel shows the full (backbone atom) structure, while the right panel compares only the β-turn region.

7 Global optimization and torsion angle dynamics

The modified constrained global optimization was also applied to the Compstatin structure prediction problem using the same constraint function and parameters [131]. The goal of introducing TAD as a component of the upper bound solution approach is to increase the number of feasible points available for initialization of the constrained local minimization. Initially TAD is used in combination with simple van der Waals overlap restraints to drive the distance violations to zero. Taken independently, this methodology is comparable to the typical implementation of TAD for NMR structure prediction [91].

To gauge the performance of the combined αBB and TAD constrained approach, a comparison was made to an independent TAD method (DYANA [91]) for solving distance restraint problems. The same dihedral angle and 53 medium and long range distance restraints were considered, along with additional distance restraints to prevent van der Waals overlaps. The coupled simulated annealing / TAD protocol was applied to a starting sample of 1000 randomly generated structures, from which a subset consisting of 20 conformers exhibiting the best target values were used to initialize a second set of runs. The five conformations with the best target function values were selected for further analysis, including initialization for constrained local minimizations with $E_{\mathrm{ECEPP/3}}$. The DYANA conformers satisfy the corresponding constraint, although their energy values are more than 70 kcal/mol higher than that of the global minimum energy structure. An analysis of the structural characteristics also indicates that the type I β-turn does not form along the Gln^5-Gly^8 b backbone for these structures. These results reflect the potential deficiencies of the independent TAD algorithm; that is, the simplified force field term is insufficient for sparse sets of distance restraints.

The use of TAD in the context of the global optimization approach surmounts this difficulty by using an iterative TAD scheme with two forms of the target function. The first set of TAD runs focuses on the reduction of the distance violations, while employing a simplified forcefield in the form of additional distance restraints to avoid atomic overlaps. This approach mimics the effects of a typical TAD approach for structure prediction. To ensure that these conformers provide low energy, this step is then followed by unconstrained minimization with a hybrid distance and ECEPP/3 energy objective function. If the ECEPP/3 energy is acceptably low, the algorithm proceeds to the constrained local minimization step, otherwise an iterative set of TAD runs are performed with readjustment of the relative weight of the distance and ECEPP/3 terms. Figure 32 shows a typical sequence for both the ECEPP/3 and distance violations energy during one solution of the upper bounding problem for Compstatin.

The results of the combined constrained global optimization and TAD algorithm can be assessed by examining the sequence of ECEPP/3 energies obtained from the solution of the upper bounding problems, as depicted in Figure 33. When compared to the original algorithm, the TAD implementation augments the number of feasible starting points by more than a factor of two. This enhancement leads to earlier identification of low energy conformers. In particular, conformers with energies less than -70 kcal/mol, and thus lower in energy than the locally minimized PDB structures, are identified within the first 10 iterations of the global optimization approach. This

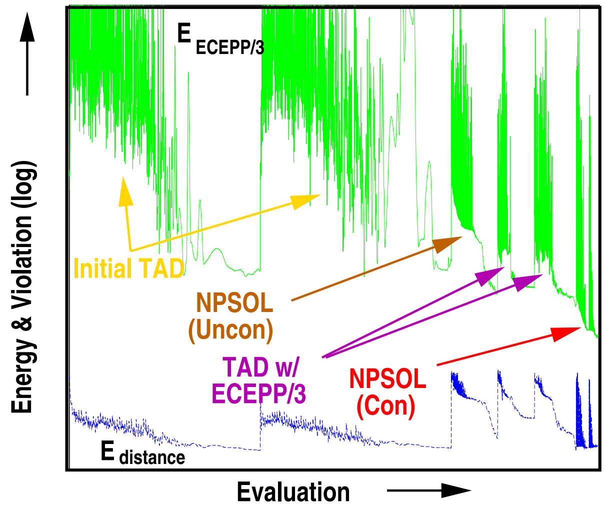

Figure 32: Log plot of $E_{\mathrm{ECEPP/3}}$ and $E^{distance}$ during a typical solution to the upper bounding problem for C3.

property has important algorithmic implications, including the ability to fathom regions based on the current estimate of the global minimum. In general, the TAD enhanced search provides more consistent and denser population of low energy conformers.

Both experimental and theoretical methods exist for the prediction of protein structures. In both cases, additional restraints on the molecular system can be derived and used to formulate a nonconvex optimization problem. Here, the traditional unconstrained problem was recast as a constrained global optimization problem, and applied to protein structure prediction using NMR data. Both the formulation and solution approach of this method differ from traditional techniques, which generally rely on the optimization of penalty-type target function using SA/MD protocols.

As a first step, the penalty type restraint functions were replaced by nonlinear constraints, which can be individually enumerated for all or subsets of the distance restraints. In addition, the objective function was transformed to a full atom force field potential, a modification that should be particularly useful for systems possessing sparse set of restraints. To solve this reformulated molecular structure prediction problem the concepts of a deterministic global optimization approach, αBB, were applied. This methodology can be used to develop theoretical guarantees for convergence to the global minimum of nonconvex constrained problems. The algorithm was further enhanced by modifying the upper bounding solution approach to include an iterative scheme involving TAD.

The approach was applied to the Compstatin structure prediction problem using both the original TAD approach and the coupled αBB-TAD approach. When considering basic structural features, such as the formation of a type I β-turn, the predicted structure was found to agree with results based on X-PLOR [32]. However, constrained global optimization was able to identify conformers with significantly lower energies than those obtained from either local minimization or

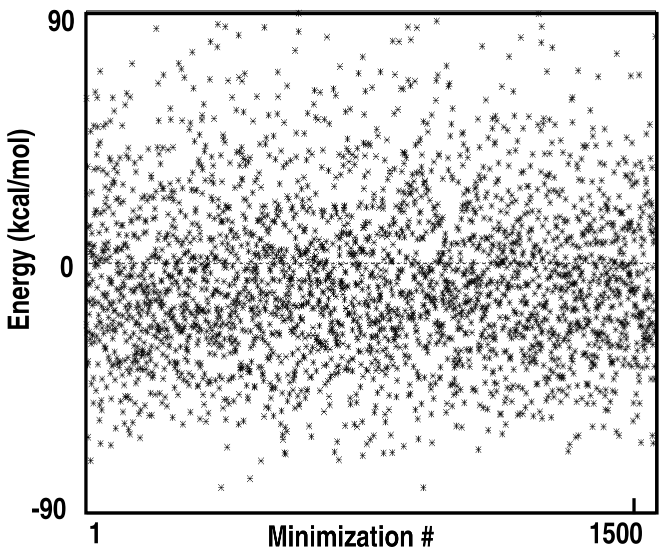

Figure 33: Energy values for Compstatin conformers obtained from combined constrained global optimization and TAD algorithm.

independent TAD algorithms. In particular, the coupled αBB-TAD implementation consistently produced dense populations of low energy conformers.

C Perspectives and Future Work

1 Structure prediction of polypeptides

In spite of pioneering contributions and decades of effort, the ab initio prediction of the folded structure of a protein remains a very challenging problem. The approaches for the structure prediction of polypeptides can be classified as : (i) homology or comparative modeling methods, (ii) fold recognition or threading methods, (iii) ab initio methods that utilize knowledge-based information from structural databases (e.g., secondary and/or tertiary structure restraints), and (iv) ab initio methods without the aid of knowledge-based information.

Knowledge-based ab initio methods exploit information available from protein databases regarding secondary structure, introduce distance constraints, and extract similar fragments from multiple sequence alignments in an attempt to simplify the prediction of the folded three-dimensional protein structure. Significant contributions include the work of Levitt and co-workers, [149, 98], Skolnick and co-workers, [187, 218], Baker and co-workers, [217, 214], Dill and co-workers, [225], and Friesner and co-workers, [174, 175, 220]. Ab initio methods that are not guided by knowledge-based information represent the most challenging category. Important advances include the pioneering work of Scheraga and co-workers [206, 153, 145], Rose and co-workers [219], and Dill and co-workers [252, 60]. Orengo et al. (1999) [186] provide a recent assessment of the current status of both types of ab initio protein structure prediction approaches.

We have recently developed the novel ASTRO-FOLD approach for the *ab initio* prediction of the three dimensional structures of proteins [130]. The four stages of the approach are outlined in Figure 34. The first stage involves the identification of helical segments and is accomplished by partitioning the aminoacid sequence into pentapeptides such that consecutive pentapeptides posses an overlap of four aminoacids; atomistic level modeling using the selected force field; generating an ensemble of low energy conformations; calculating free energies that include entropic, cavity formation, polarization and ionization contributions for each pentapeptide; and calculating helix propensities for each residue using equilibrium occupational probabilities of helical clusters.

In the second stage, β strands, β sheets, and disulfide bridges are identified through a novel superstructure-based mathematical framework originally established for chemical process synthesis problems [70]. Two types of superstructure are introduced, both of which emanate from the principle that hydrophobic interactions drive the formation of β structure. The first one, denoted as *hydrophobic residue-based superstructure*, encompasses all potential contacts between pairs of hydrophobic residues (i.e., a contact between two hydrophobic residues may or may not exist) that are not contained in helices (except cystines which are allowed to have cystine-cystine contacts even though they may be in helices). The second one, denoted as β *strand-based superstructure*, includes all possible β strand arrangements of interest (i.e., a β strand may or may not exist) in addition to the potential contacts between hydrophobic residues. The hydrophobic residue-based and β strand-based superstructures are formulated as mathematical models which feature three types of binary variables : (i) representing the existence or nonexistence of contacts between pairs of hydrophobic residues; (ii) denoting the existence or nonexistence of the postulated β strands; and (iii) representing the potential connectivity of the postulated β strands. Several sets of constraints in the model enforce physically legitimate configurations for antiparallel or parallel β strands and disulfide bridges, while the objective function maximizes the total hydrophobic contact energy. The resulting mathematical models are Integer Linear Programming (ILP) problems which not only can be solved to global optimality, but can also provide a rank ordered list of alternate β sheet configurations.

The third stage serves as a preparative phase for atomistic-level tertiary structure prediction, and therefore focuses on the determination of pertinent information from the results of the previous two stages. This involves the introduction of lower and upper bounds on dihedral angles of residues belonging to predicted helices or β strands, as well as restraints between the C^{α} atoms for residues of the selected β sheet and disulfide bridge configuration. Furthermore, for segments which are not classified as helices or β strands, free energy runs of overlapping heptapeptides are conducted to identify tighter bounds on their dihedral angles.

The fourth and final stage of the approach involves the prediction of the tertiary structure of the full protein sequence. The problem formulation, which relies on dihedral angle and atomic distance restraints acquired from the previous stage, is equivalent to the problem outlined in Section III.B. The generation of low energy starting points for constrained minimization is enhanced by introducing torsion angle dynamics [91] within the context of the αBB global optimization

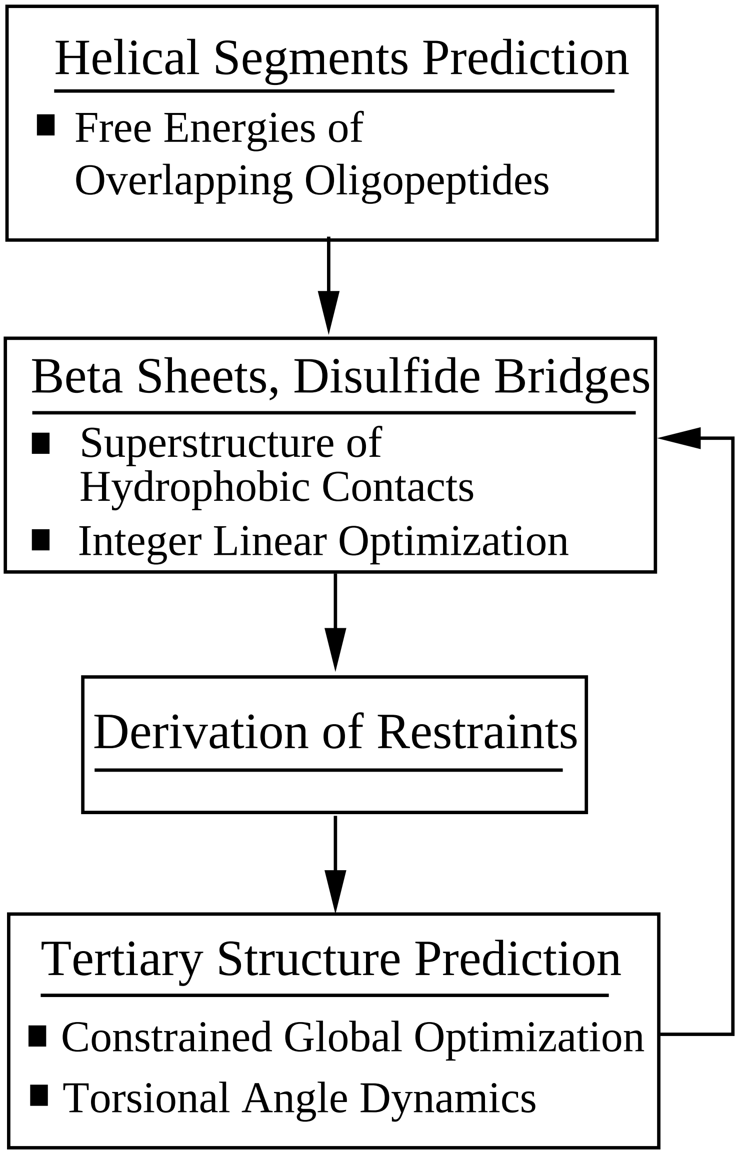

Figure 34: Overall flowchart for the ab initio structure prediction using ASTRO-FOLD. The first stage addresses the prediction of helical segments based on free energy calculations of overlapping oligopeptides. The second stage introduces a superstructure-based framework coupled with integer-linear optimization for the prediction of a rank-ordered list of β sheets and disulfide bridges. The third stage derives lower and upper bounds on the (ϕ,ψ) dihedral angles of the secondary structure residues, the distances between pairs of contacts of hydrophobic residues, and the (ϕ,ψ) angles of the loop/turn.residues. The fourth stage introduces a constrained formulation for the tertiary structure prediction and its solution via the αBB global optimization approach enhanced by torsion angle dynamics. An iterative loop over the final three stages allows for analysis of multiple β sheet and disulfide bridge configurations.

framework, as described in Section III.B.7.

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 rapid 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, and suggests that tertiary and secondary features form concurrently. This work bridges the gap between the two viewpoints by introducing a novel ab initio approach for tertiary structure prediction in which helix nucleation is controlled by local interactions, while non local hydrophobic forces drive the formation of β structure. The agreement between the experimental and predicted structures validates the use of the ASTRO-FOLD method for generic tertiary structure prediction of polypeptides.

2 Parallelization issues

The extension of our global optimization approaches to larger protein systems requires the use of distributed computing environments. Such implementations have been developed independently of system architecture, and the code has been compiled and optimized using the MPI (message passing interface) standard.

On a fundamental level, these parallel implementations exploit the inherent branch-and-bound structure of the αBB algorithm. A major 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 result in better approximations. This is equivalent to simultaneously exploring multiple domains in order to perform a more efficient search, which is the rationale behind advocating the development of a parallel algorithm.

Distributed frameworks for branch-and-bound algorithms can rely on two basic protocols. The most simplistic 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.

Initial implementations of the αBB algorithm have employed the tree hierarchy through a master-slave decomposition approach. This 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 can also encompass several levels in the branch and bound tree depending on the computational requirements for solving one node in the tree. This procedure can be efficient for treating large protein systems because of low communication time overhead. 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.

The overall protein folding solution approach also affords other levels of parallelism. For example during the helix prediction phase, the full protein is decomposed into smaller segments. This decomposition allows us to identify the major secondary structural components (α helical, β-sheet) of the protein by solving smaller global optimization subproblems (using αBB) in parallel. The extent of parallelism depends on the length of these subsegments and the parallelism of the underlying αBB algorithm.

IV Dynamics of Protein Folding

A Background

The protein folding problem is a very important problem in computational chemistry and molecular biology. The ability of a protein to function properly within the cell depends on its tertiary structure. Considering how precisely and reliably a protein shapes itself to perform its specific task, very little is understood about the mechanism of protein folding. Better understanding and insight on the mechanism of protein folding are of major importance.

In Section III, we discussed the structure prediction problem, in which the native conformation is sought. In this section, we pursue the protein-folding problem further by studying the *folding mechanism*, that is, the pathways followed by a protein as it proceeds from its initial (extended) conformation to its native state, as well as the rates associated with these folding processes.

1 Studying the Dynamics of Secondary Structure Formation

According to the hierarchic model of protein folding, the time scale of formation of secondary structures, such as alpha-helices and beta-sheets, within a given protein occurs on a much shorter time scale than the formation of tertiary structure. Whether this is true or not, there is much evidence — both theoretical and experimental — that the folding of large proteins begins with the formation of these secondary structure elements [16, 17]. Therefore, an initial step in understanding protein folding is understanding the folding process of the secondary structures such as alpha-helices and beta-sheets. Insights can be gained into the folding mechanism of these structures by studying short peptides which exhibit the structure we wish to study.

Alpha helices have been studied for a relatively long time [40, 64]. Numerous short peptides have been observed in the lab to form alpha helices in solution, and have been the subject of many experimental and theoretical studies [94, 21, 50, 45, 160, 212, 247, 246]. Our recent analysis of tetra-alanine, the shortest peptide to form an alpha helix, has provided us with enormous insights into the folding mechanism of these structures, and will be presented in Section IV.C.

The situation is very different for beta-sheet structures. Until recently, experimental studies of these structures have been mostly unsuccessful, largely due to the fact that short peptides which fold into a beta-sheet conformation tend to aggregate in solution [64]. These difficulties have finally been overcome with the recent discoveries of designed sequences, such as Beta-nova

[135] and others [193, 54], as well as the second beta-hairpin fragment of Protein G (residues 41–56) [26], thus opening the door to a proper study of beta-hairpin and beta-sheet formation [210, 189, 61, 100, 154, 178, 65, 34, 30]. Our ongoing efforts to analyze the Protein G fragment (41–56) will be discussed in Section IV.E.

2 Searching for Stationary Points

A promising approach to understanding protein folding is the study of its potential energy surface. The first step in the study of any potential energy surface is the identification of stationary points (local minima and saddle points), since these points play a crucial role in defining the topography of the surface. The local minima represent stable configurations of the protein molecule, and the first-order saddle points generally correspond to transition states which connect two such configurations. A protein-folding process can be thought of as a transition between two local minima through a transition state, or a series of such transitions.

The problem of finding stationary points of a potential energy surface is an old one, and numerous methods have been developed to solve it. The most obvious method is applying the Newton-Raphson method to the equation $\nabla V = 0$. The Newton-Raphson method tends to yield a solution whenever the initial guess is close to a stationary point and the Hessian matrix has the appropriate signature for the type of stationary point desired (minima, first-order saddle, etc). It cannot be used, for example, to walk away from a local minimum towards a first-order saddle point.

The various "eigenmode-following" methods are sophisticated variants of the Newton-Raphson method [228, 39, 216, 20, 184, 49]. The Hessian is diagonalized, and a modified Newton-Raphson step is generated by "shifting" some of the eigenvalues of the Hessian, from positive to negative or vice versa, before applying its inverse. These methods allow one to step away from local minima in search of transition states, and vice versa.

There are a number of stochastic methods used to find stationary points [24]. Local minima can be obtained by frequent quenching of a constant energy (or temperature) trajectory [222]. Simulated annealing by running a constant temperature trajectory simulation, slowly reducing the temperature to zero in the process, can sometimes lead to good candidates for the global minimum. The method of "slowest slides" [25] has been used to search for transition states connecting two given local minima: a constant energy trajectory is followed during a transition from one local minimum to the other, and the maximum along that trajectory is taken as an initial guess for the transition state.

The global minimum can also be found by use of genetic algorithms, in which new conformations are generated from old conformations by random mutations in the hope of eventually lowering the potential energy. Of particular interest to us is the Conformation Space Annealing (CSA) algorithm [144, 145, 143], which is a combination of genetic, annealing, and buildup methods. This algorithm can also be used to generate a variety of low energy conformations.

Other methods exist for searching for the global minimum of a potential energy surface. Diffusion equation and distance scaling methods have been applied to the problem of finding the global minimum of a potential energy surface [242]. Smoothing transformations are applied to the potential energy surface to eliminate the irrelevant local minima. The remaining minima are tracked back to the original potential energy surface as the transformations are gradually removed. Another method involves obtaining a large sample of local minima and forming a "convex global underestimator" of the potential energy surface based on those sample points [59]. The global minimum of the original potential energy surface is sought in the vicinity of the global minimum of the convex global underestimator.

Many dynamical studies of protein folding are carried out these days by performing molecular dynamics simulations, in which the time-evolution of the protein's configuration is determined directly by solving Newton's equations of motion. Not only is it possible to obtain numerous low-energy minima in the vicinity of the starting point, but rate and pathway information can also be inferred directly from the trajectories generated by these simulations. A major drawback of these simulations, however, is their computational expense. Current limitations on simulations are on the order of a few hundred nanoseconds of real time (a $1\,\mu$s simulation has been reported recently [62]), which is far too short to enable a full simulation of the folding process of even a modest sized protein.

All of these methods, good in their own right, share one very important drawback: there is no guarantee that all (or even the most important) local minima and first or higher-order transition states will be found. In this article, we propose a method of finding all stationary points of a given potential energy surface in which we apply the αBB deterministic branch and bound global optimization algorithm to the system of equations $\partial V/\partial x_i = 0$. The general algorithm is discussed in II.B, and its specific application to the stationary point search is discussed in IV.B. We have successfully applied this method to small systems, such as triatomic molecules, alanine, alanine dipeptide, and tetra-alanine [247, 246]. We will discuss tetra-alanine in Section IV.C.

3 Analyzing the Potential Energy Surface

Once the minima and first-order saddles are determined, the potential energy surface can be analyzed. The folding mechanism of the protein can be understood by enumerating the reaction pathways from the extended conformations to the native state. The first step in constructing the pathways is to determine for each transition state which two minima it connects. This is accomplished by performing a downhill search from the transition state along each of the two reaction coordinate directions. The result is a list of minimum–saddle–minimum "triples". The reaction pathways can then be enumerated by joining these triples together in chains using graph theory techniques.

Transition rates can be calculated using Rice–Ramsperger–Kassel–Marcus (RRKM) theory [110]. The basic assumptions of RRKM theory is that the protein can be treated thermodynamically in the vicinity of the minima as well as the transition state, and that the transition is completed once the transition state is crossed (i.e., there are no re-crossings). Once the transition rates have been determined, the Master equation can be solved for the occupation probabilities of each state as functions of time. This gives us a direct indication of how long it takes for a protein prepared in a given unfolded state to reach its native state. It is also possible to use this information to calculate the time-evolution of other quantities, such as (ensemble-averaged) energies, atomic distances, and dihedral angles.

Becker and Karplus proposed a graphical representation of the topography of a potential energy surface [21] based on the connectivity tree originally introduced by Czerminiski and Elber [50]. They define a finite energy (temperature) generalization of the "catchment region". As the energy (temperature) is increased, regions that were once disconnected by high barriers begin to merge. This coalescence process is described by means of a "energy (temperature) disconnectivity graph". The shape of the disconnectivity graph reveals an enormous wealth of dynamical information. We extended this idea by constructing a "rate disconnectivity graph" which is based on transition rates, rather than energy levels or barrier heights.

We have applied these methods to tetra-alanine (an alpha-helix), which we discuss in Section IV.C, and the 41–56 fragment of Protein G (a beta-hairpin), which we discuss in Section IV.E.

B The αBB Global Optimization Approach

Stationary points of all orders (i.e., minima, maxima, first order and higher order transition states) of a given potential energy surface $V(\mathbf{x})$ are determined by the constraints

$$\frac{\partial V}{\partial x_i} = 0 \ , \qquad i = 1, \ldots, N_x \tag{88}$$

where N_x is the number of variables: $\mathbf{x} = (x_1, \ldots, x_{N_x})$. Equation (88) is an example of a nonlinearly constrained system of algebraic equations. Indeed, (88) can be obtained from (17) in Section II.B.1 by assigning $f_i(\mathbf{x}) = \partial V/\partial x_i$ for $i = 1, \ldots, N_f = N_x$, and $N_g = 0$.

In Section II.B, we explained how such systems of equations can be solved using the αBB global optimization algorithm. This algorithm applies whenever the constraint functions $\partial V/\partial x_i$ are twice continuously differentiable (C^2), or in other words, whenever the potential energy function itself is C^3. Unlike other methods of locating stationary points, the αBB provides a rigorous theoretical guarantee of finding *all* of the stationary points on a potential energy surface.

According to the αBB algorithm, the original problem (88) is first re-expressed as a global optimization problem by introducing a slack variable:

$$\begin{aligned} & \min_{\mathbf{x},s} s \\ \text{subject to} \qquad & \partial V/\partial x_i - s \le 0 \ , \qquad i = 1, \ldots, N_x \\ & -\partial V/\partial x_i - s \le 0 \ , \qquad i = 1, \ldots, N_x \\ & \mathbf{x}^L \le \mathbf{x} \le \mathbf{x}^U \end{aligned} \tag{89}$$

The global minima of (89) with $s = 0$ correspond to solutions to the original problem (88).

Configuration space is searched for stationary points by subdividing the full conformational space into smaller and smaller regions. At each stage, the current region is tested for possible stationary points by solving the lower bounding problem:

$$\min_{\mathbf{x},s} s$$

$$
\begin{aligned}
\text{subject to} \qquad & \partial V/\partial x_i - \alpha_i^+ \sum_k (x_k^U - x_k)(x_k - x_k^L) - s \le 0 \\
& -\partial V/\partial x_i - \alpha_i^- \sum_k (x_k^U - x_k)(x_k - x_k^L) - s \le 0 \\
& \mathbf{x}^L \le \mathbf{x} \le \mathbf{x}^U
\end{aligned} \tag{90}
$$

The LHS of each constraint in (90) is a convex underestimator of the corresponding term in (89), and is obtained by subtracting off a sufficiently large quadratic term. The lower bounding problem (90) is indeed convex, provided that the coefficients $\alpha_i^{\pm}$ satisfy

$$
\begin{aligned}
\alpha_i^+ &\ge -\frac{1}{2} \min_{\mathbf{x} \in [\mathbf{x}^L, \mathbf{x}^U]} \{\lambda_k(H_{\partial V/\partial x_i}(\mathbf{x})), 0\} \\
\alpha_i^- &\ge +\frac{1}{2} \max_{\mathbf{x} \in [\mathbf{x}^L, \mathbf{x}^U]} \{\lambda_k(H_{\partial V/\partial x_i}(\mathbf{x})), 0\}
\end{aligned} \tag{91}
$$

Assuming that (91) is satisfied, the lower bounding problem is convex and can be solved to global optimality by any commercial local optimization package. The global minimum s_{LB} of (90) provides a valid lower bound of the global minimum of (89), and thus can be used to check if a stationary point can exist in the current region $[\mathbf{x}^L, \mathbf{x}^U]$. If $s_{\mathrm{LB}} > 0$, no such solution exists, and the region can be fathomed. If $s_{\mathrm{LB}} \le 0$, then a solution may or may not exist in $[\mathbf{x}^L, \mathbf{x}^U]$, and so that region will be subdivided and both subregions checked by the same procedure. The αBB algorithm terminates when all regions have either been fathomed, or reduced sufficiently in size at which point a solution to (88) is obtained by a local search.

Calculating values of $\alpha_i^{\pm}$ according to (91) is difficult in general because the Hessian matrices $H_{\partial V/\partial x_i}$ depend on $\mathbf{x}$. A simplified method of calculating $\alpha_i^{\pm}$ is to start with small values of $\alpha_i^{\pm}$ (e.g., $\alpha_i^{\pm} = 5$) and increase the values of $\alpha_i^{\pm}$ until no new solutions are found. This can be a practical solution to many problems where the correct values of $\alpha_i^{\pm}$ are difficult to determine. However, this method has the one serious drawback in that it sacrifices the theoretical guarantee of finding *all* solutions. In spite of this fact, we were able to identify all minima and first-order transition states using modest values of $\alpha_i^{\pm}$ for alanine, alanine dipeptide, and tetra-alanine. Tetra-alanine will be discussed in Section IV.C.

A more robust method involves calculating the Hessian matrices $H_{\partial V/\partial x_i}$ at various grid points to get a sample of required $\alpha_i^{\pm}$ values. First we select a grid, $\{\mathbf{x}^k\}$. Then we evaluate the Hessian for each constraint at each grid point, $H_{\partial V/\partial x_i}(\mathbf{x}^k)$, and use (91) to determine precomputed values of $\alpha_i^{\pm}(\mathbf{x}^k)$ at each grid-point. During the αBB run, appropriate values of $\alpha_i^{\pm}$ for a given region are determined by selecting the maximum $\alpha_i^{\pm}$ over all grid-points contained in the region. This method of generating $\alpha_i^{\pm}$ was used when we studied triatomic molecules, which is discussed in [247].

C Dynamics of Coil to Helix Transitions

In this section, we attempt to elucidate the formation of alpha-helices by studying tetra-alanine, which is one of the smallest peptides that can exhibit a full alpha-helical turn. Tetra-alanine is depicted in Figure 35.

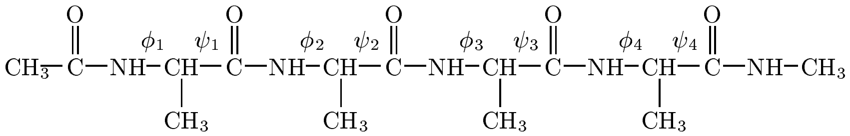

Figure 35: Tetra-alanine

Table 27: Eigenmode III results for unsolvated tetra-alanine.

	4^8 grid	6^8 grid
Local minima	16125	62373
1st-order saddles	18902	212938

In Sections IV.C.1–IV.C.6, we study tetra-alanine in vacuum. We use the ECEPP/3 potential energy surface [180] (see Section III.A.1 and Figure 11), which is an all-atom potential energy function. In Section IV.C.7, we consider tetra-alanine in solution by adding a solvation free-energy term to the ECEPP/3 potential energy surface. The solvation free energy is modeled by the volume method using the Reduced Radius Independent Gaussian Sphere (RRIGS) approximation (see Section III.A.2). To simplify the calculations, we fix bond lengths and bond angles, allowing only the eight backbone (ϕ, ψ) dihedral angles to vary.

1 Stationary Points for Unsolvated Tetra-Alanine

The first step in elucidating the folding process of tetra-alanine is to determine the local minima and first-order saddles of its potential energy surface. We first obtained a testbed of minima and first-order saddles by applying a brute-force eigenmode-following search (Eigenmode III [228]) using a grid of starting points. Our search results are summarized in Table 27. For our initial attempt to analyze tetra-alanine [247], we generated a 4^8 grid of starting points and performed minimum and first-order saddle searches from each point. The transition states were then followed down to the minima they connect, resulting in additional minima found. Given the relative high percentage of starting points which resulted in unique stationary points, we decided to increase the grid to 6^8 and perform first-order saddle searches from each point. Additional minima were obtained by following each such transition state down to the minima they connect. After merging these new results with the results from the 4^8 grid, we had generated a total of 62373 minima and 212938 first-order saddle points [246].

Tetra-alanine is one of the smallest peptides that can exhibit an alpha-helical conformation as well as an extended conformation. These two conformation types can characterized by their (ϕ, ψ) angle values. Alpha helical conformations tend to have (ϕ, ψ) angle values in the vicinity of $(300°, 300°)$. On the other hand, extended conformations tend to have (ϕ, ψ) values in the vicinity of $(300°, 120°)$.

Therefore, to facilitate the classification of tetra-alanine conformations, we subdivide the (ϕ, ψ) plane into regions and classify those regions according to Table 28. Values of (ϕ, ψ) corresponding

Table 28: Classification scheme for (ϕ, ψ) pair.

Symbol	ψ	Decoration	ϕ
a	$270° \leq \psi \leq 335°$	No prime	$270° \leq \phi \leq 330°$
i	$335° \leq \psi$ or $\psi \leq 90°$	Prime	$180° \leq \phi \leq 270°$
b	$90° \leq \psi \leq 150°$	Double prime	otherwise
j	$150° \leq \psi \leq 270°$		

Table 29: Ground state and extended conformation of unsolvated tetra-alanine.

Minimum	Classification	E (kcal/mole)	F (kcal/mole)
min.1	aaaa	−6.643	−11.798
min.1587	bbbb	4.916	−5.549

to alpha-helix formation are classified as "a", and values of (ϕ, ψ) corresponding to beta-sheet formation are classified as "b". Each conformation of tetra-alanine is characterized by four (ϕ, ψ) pairs, and hence can be classified by a concatenation of four symbols.

Of the 62373 minima, we found one alpha-helical conformation, min.1 (aaaa), and one extended conformation, min.1587 (bbbb). Their potential energy and free energy[1] values can be found in Table 29. The alpha-helix conformation is the lowest energy conformation of tetra-alanine. We will be concentrating on the folding process from the extended conformation to the ground state.

We checked the αBB algorithm described in Section IV.B against the Eigenmode III search for stationary points by conducting αBB runs on selected regions of the potential energy surface. Selected results are given in Table 30. We started with a constant value of $\alpha = 20$, and then increased α in subsequent runs until we found all stationary points located by the Eigenmode III search. In all cases, modest values of α (less than 100) were sufficient to locate all minima and first-order saddles found by Eigenmode III. In many cases, additional saddle points were located.

2 Transition Rates and the Master Equation

Having now identified the local minima and first-order transition states, we are now in a position to enumerate the reaction pathways between states and calculate transition rates. The connectivity between the various minima is determined by following each transition state back to the minima they connect. This is accomplished by perturbing the transition state in each of the two directions along the reaction coordinate, and then using Eigenmode III to locate a local minimum from that starting point. This gives us a list of (minimum, transition state, minimum) triples.

We can then calculate the transition rate matrix using Rice-Ramsperger-Kassel-Marcus (RRKM) theory. According to RRKM theory [110, 137, 247], the transition rate for a single transition is

[1]By "free energy", we mean potential energy plus the contributions from vibrational entropy. Free energy can be calculated using (93) in Section IV.C.2.

Table 30: Selected results from αBB tetra-alanine runs.

Region	Saddle Type	Eigenmode III	αBB	α
aaaa	min	1	1	25
bbbb	min	1	1	20
	1st	4	4	
	2nd	6	6	
	3rd	4	4	
	4th	1	1	
bibi	min	1	1	20
	1st	1	2	
	2nd	0	1	
bbbj′	min	2	2	20
	1st	8	9	
	2nd	4	17	
	3rd	3	16	
	4th	2	7	
	5th	0	1	
aai′i	min	2	2	80
	1st	1	1	

given by

$$W_{j' \to \text{ts} \to j} = \frac{kT}{h} \frac{Q_{\text{ts}}}{Q_{j'}} \tag{92}$$

The partition functions at the minima and first-order saddles are related to the free-energies of those stationary points, and can be evaluated using the harmonic approximation

$$Q = e^{-F/kT} = e^{-E/kT} \prod_i \frac{kT}{h\nu_i} \tag{93}$$

where E and F are the potential energy and free energy, respectively, of the stationary point, and ν_i are the vibrational frequencies of the molecule around the stationary point. The product over frequencies takes into account the vibrational entropy of the system. Substituting (93) into (92) yields

$$W_{j' \to \text{ts} \to j} = \frac{\prod_i \nu_i^{j'}}{\prod_{i \neq \text{r.c.}} \nu_i^{\text{ts}}} e^{-(E_{\text{ts}} - E_{j'})/kT}$$

Summing over all transition states connecting two particular minima yields the transition rate matrix

$$W_{jj'} = \sum_{\text{ts}} W_{j' \to \text{ts} \to j} \tag{94}$$

The time-evolution of occupation probabilities can be calculated by solving the Master equation

$$\frac{dP_j}{dt} = w_{jj'} P_{j'}(t)$$

where

$$w_{jj'} = \begin{cases} W_{jj'} & \text{if } j \neq j' \\ -\sum_{j''} W_{j''j} & \text{if } j = j' \end{cases}$$

Coupled differential equations like (94) are solved by diagonalizing the matrix $w_{jj'}$, so that

$$\sum_{j'} w_{jj'} u_{j'}^{(i)} = \lambda^{(i)} u_j^{(i)}$$

The general solution to (94) can be written in the form

$$P_j(t) = \sum_i a_i e^{\lambda^{(i)} t} u_j^{(i)} \tag{95}$$

where the coefficients a_i are determined by the initial probability distribution at $t = 0$.

One of the eigenvalues $\lambda^{(0)}$ is zero. The associated eigenvector corresponds to the equilibrium ($t = \infty$) probability distribution,

$$u_j^{(0)} = P_j(+\infty) = Q_j / \sum_{j'} Q_{j'}$$

All other eigenvalues are negative, and correspond to transient probabilities with a decay time of $\tau^{(i)} = -1/\lambda^{(i)}$.

The time-evolution of occupation probabilities for the extended conformation and the 3 lowest free energy states of unsolvated tetra-alanine at room temperature $T = 300\,\mathrm{K}$, starting with the extended conformation at $t = 0$ (i.e., $P_{\mathrm{bbbb}}(0) = 1$, all other $P_j(0) = 0$) is given in Figure 36. It takes tetra-alanine about 10^{-10} sec to reach the ground state from the extended conformation.

3 Pathways

Details of the folding process can be determined by enumerating the pathways from the extended conformation to the ground state. A pathway is defined as a sequence of minima joined together by transition states.

$$\text{initial state} \rightarrow \text{ts} \rightarrow \text{min} \rightarrow \text{ts} \rightarrow \text{min} \rightarrow \ldots \rightarrow \text{min} \rightarrow \text{ts} \rightarrow \text{final state}$$

Pathways between these two states can be enumerated using graph-theory techniques. We construct a graph where each node in the graph represents a minimum and each edge in the graph represents a transition state which connects two minima. The set of all pathways from one minimum to another can be generated by an exhaustive search.

If we conduct this exhaustive search without restriction, we would generate an enormous number of pathways. It is important to restrict the pathways we generate in a sensible manner. We selected pathways based on two criteria: (1) we restrict the length of the pathway (i.e., number of minima) to be less than or equal to some prescribed maximum length, and (2) we also apply a transition rate cutoff, effectively ignoring transitions whose rates fall below the cutoff value. The number of pathways from the extended conformation to the ground state of unsolvated tetra-alanine at $T = 300\,\mathrm{K}$ for various length and rate cutoffs is given in Table 31. The total number of minima and transition states involved in such pathways are given in Table 32.

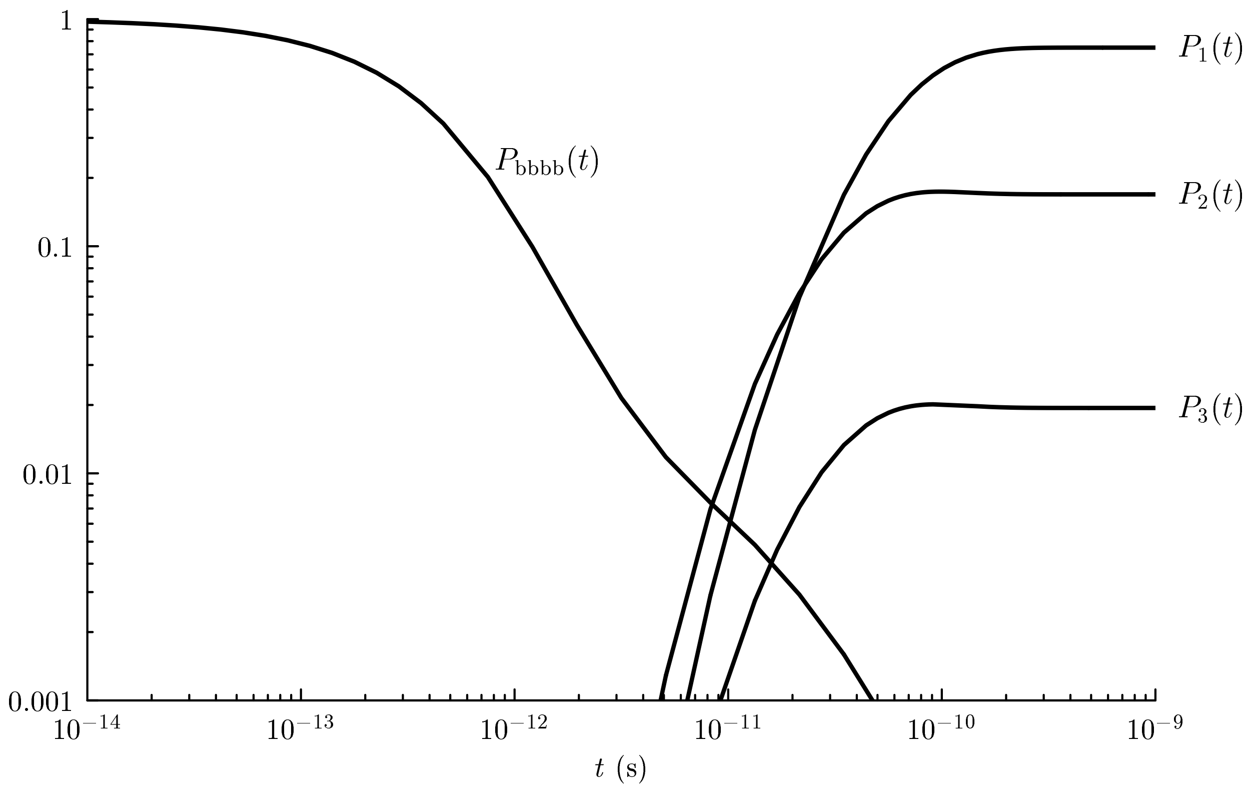

Figure 36: Time evolution of the extended conformation and the 3 lowest free energy states of unsolvated tetra-alanine at $T = 300\,\mathrm{K}$.

Table 31: Number of pathways from extended conformation to ground state with given length restriction and rate cutoff.

Maximum length	No rate cutoff	10^6 Hz	10^7 Hz	10^8 Hz	10^9 Hz	10^{10} Hz	10^{11} Hz
6							
7	4						
8	38						
9	999	421	421	421	421	285	130
10	19963	10836	10828	10828	10733	7443	2099
11	297974	150831	150396	149391	146493	92216	21004
12	4132256	1868821	1859469	1832692	1768736	1002874	221592

Table 32: Number of minima / transition states involved in pathways from the extended conformation to ground state with given length restriction and rate cutoff.

Maximum length	No rate cutoff	10^6 Hz	10^7 Hz	10^8 Hz	10^9 Hz	10^{10} Hz	10^{11} Hz
6							
7	12/ 14						
8	26/ 42						
9	236/ 488	96/ 183	96/ 183	96/ 183	96/ 183	86/ 160	65/ 114
10	886/ 2339	339/ 952	339/ 951	339/ 951	332/ 932	287/ 790	188/ 466
11	2817/ 8341	664/2177	663/2173	657/2152	651/2120	526/1696	357/1044
12	6403/21316	943/3405	938/3388	922/3341	913/3291	754/2622	509/1699

These two criteria were applied in an attempt to find the most relevant pathways. Since the faster pathways are likely to be the most important ones, it makes sense to eliminate pathways which involve one or more slow transitions (i.e., transitions which fail to meet the rate cutoff). The length cutoff is chosen for more practical reasons. Even with a transition rate cutoff, the number of pathways increases exponentially with the length cutoff (about a factor of 10 for each additional minimum). An exhaustive pathway search would be intractable if we did not impose a length cutoff. It is assumed that the fastest pathways are also among the shortest in length. Although we have no proof of this, we will see evidence later on that suggests that we have found the most relevant pathways.

We examined in detail the pathways of length 9 and 10 with a transition rate cutoff of 10^6 Hz. An example pathway of length 9 is given in Figure 37. For each such pathway, we estimated the amount of time it would take for tetra-alanine to proceed from the extended conformation to the ground state along that particular pathway by solving the Master equation for a reduced system consisting only of the minima and transition states involved in the pathway. The decay time of the longest lived transient probabilities was used as an estimate of the overall transition time. The fastest transition times were on the order of 5×10^{-11} sec, and most of the 10836 pathways we looked at had transition times less than 1×10^{-9} sec. Clearly there is no single most important pathway: there are many pathways which are all equally important. We also found that the pathways of length 9 tended to be among the fastest of the pathways of length 10 or less, suggesting that shorter pathways tend to be faster.

We also studied the pathways of length 10 or less in terms of changes in the ϕ and ψ angles. Each (ϕ, ψ) pair is classified according to Table 28. In proceeding from the extended conformation to the ground state, each of the four (ϕ, ψ) pairs must proceed from "b" to "a". We observed that this process tends to follow regular patterns.

We make the following general observations regarding the rotation of the ψ angles:

1. Each ψ angle normally progresses in the sequence b $\rightarrow$ i $\rightarrow$ a or b $\rightarrow$ j $\rightarrow$ i $\rightarrow$ a.

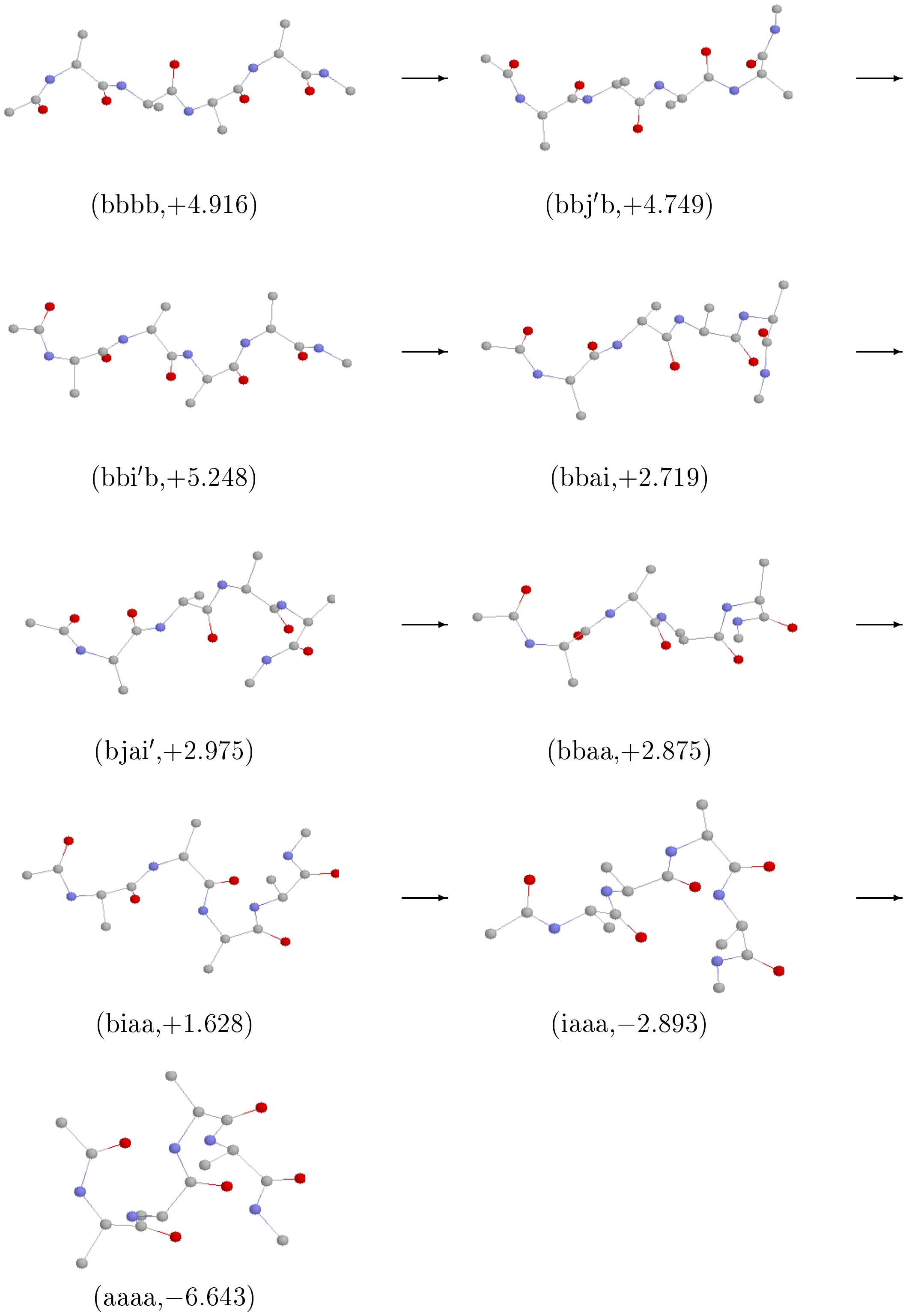

Figure 37: One possible pathway from the extended conformation to the ground state of unsolvated tetra-alanine.

2. No direct b $\rightarrow$ a transitions are observed,[2] indicating that a rotation of ψ from beta-sheet to alpha-helical values is too large for a single transition.

3. Most pathways of length 10 or less involve at least one transition where more than one ψ angle changes (cooperative motion).

4. A wide variety of cooperative motion is possible, but the two most common types are as follows:

bi $\rightarrow$ ia	36%
bj $\rightarrow$ ii	14%

5. There is a tendency for one half of the molecule to fold (nearly) completely followed by the other half (e.g., bbbb $\rightarrow$ bbaa $\rightarrow$ aaaa).

We can analyze the pathway given in Figure 37 in terms of these observations. The individual ψ angles proceed as follows:

$$\begin{aligned} \psi_1 &: \mathrm{b} \rightarrow \mathrm{i} \rightarrow \mathrm{a} \\ \psi_2 &: \mathrm{b} \rightarrow \mathrm{j} \rightarrow \mathrm{b} \rightarrow \mathrm{i} \rightarrow \mathrm{a} \\ \psi_3 &: \mathrm{b} \rightarrow \mathrm{j} \rightarrow \mathrm{i} \rightarrow \mathrm{a} \\ \psi_4 &: \mathrm{b} \rightarrow \mathrm{i} \rightarrow \mathrm{a} \end{aligned}$$

Except for a slight backtrack in ψ_2, this pathway is consistent with (1) and (2). This pathway also exhibits 3 transitions which involve cooperative motion. Two of them are in the form bi $\rightarrow$ ia, which is the most common form observed. The other cooperative motion, ji $\rightarrow$ ba (non-adjacent alanines), has also been observed but is not nearly as common as the two forms listed above. Finally, it should be remarked that this pathway does pass through a bbaa minimum. In other words, the right side (the carboxyl terminus) folds completely before the left side (the amino terminus) folds at all. Not all pathways follow this rule strictly, although we have found that tetra-alanine tends to fold its right side most of the way before its left side makes significant progress.

The rotation of the ϕ angles plays less of a role in the folding process than rotation of ψ angles. ϕ takes on similar values for alpha-helical and beta-sheet conformations. We found that the very slowest transitions (on the order of 100 Hz or less) tend to involve rotations of the ϕ angles from inside to outside of the range $180° \leq \phi \leq 330°$ and vice versa. In fact, none of the minima involved in pathways of any length with a rate cutoff of 10^6 Hz involves ϕ angles outside this range (they would be indicated in our classification scheme by a double-prime). This can be proved rigorously by examination of the rate disconnectivity graph, which we will discuss next.

4 Rate Disconnectivity Graph

We constructed the rate disconnectivity graph for tetra-alanine at $T = 300\,\mathrm{K}$. It is shown in Figure 38. The rate disconnectivity graph provides us with the rate-dependent connectivity of the

[2] This has been checked rigorously for all pathways length 11 or less with a rate cutoff of 10^6 Hz. What we have in fact found is that there are transition states which connect two minima b $\rightarrow$ a, but either the transition itself is very slow, or else the minima are so high in energy that it seems unlikely that a fast pathway (of *any* length) could pass through it. Our conclusion is that b $\rightarrow$ a is not observed for all but the very slow pathways.

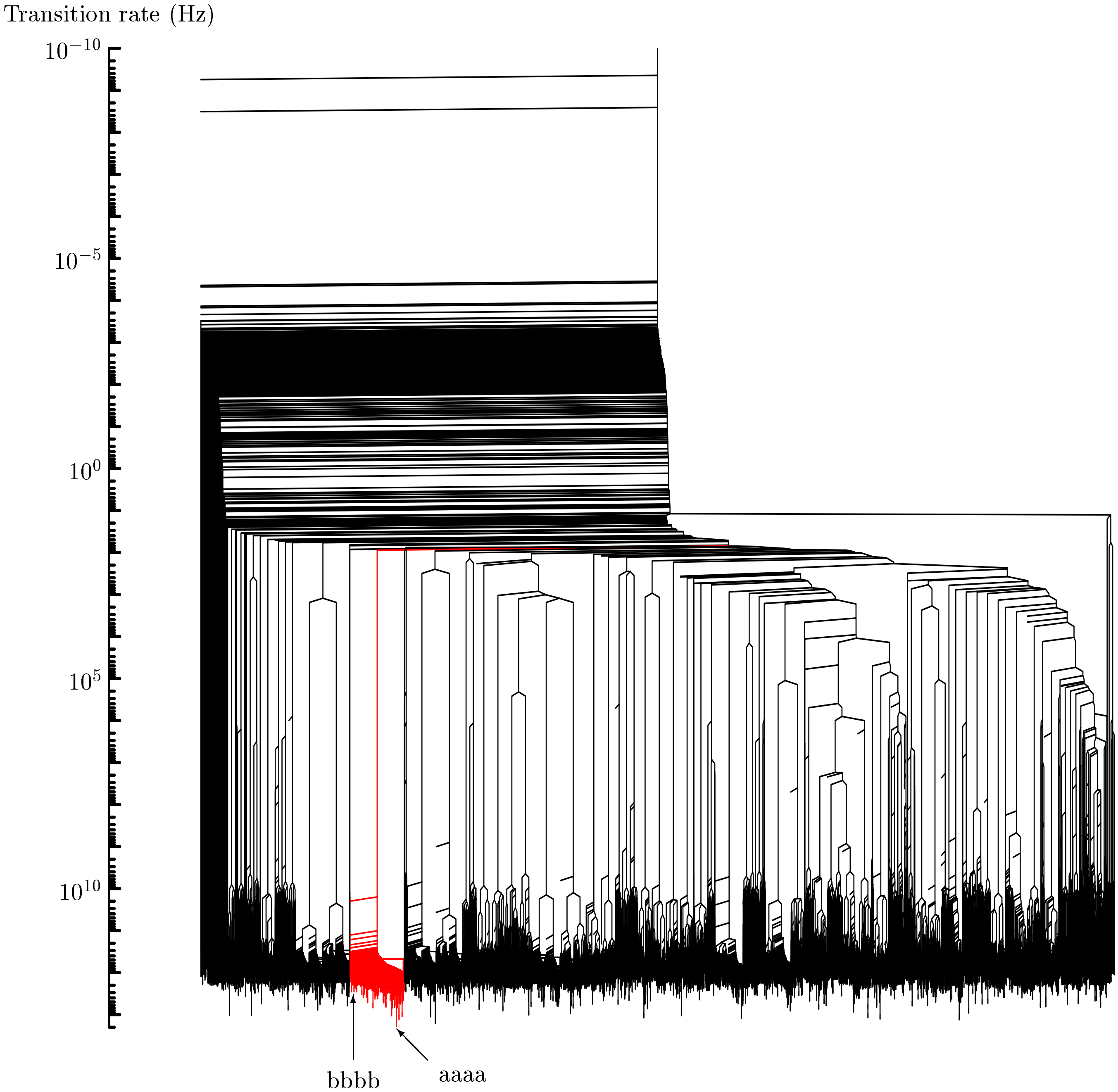

Figure 38: Complete rate disconnectivity graph for unsolvated tetra-alanine at $T = 300\,\mathrm{K}$. The alpha-helical ground state and the extended conformation both lie in the highlighted subtree.

potential energy surface [247, 246, 50, 21]. If we begin at the top of the graph, with a very small rate cutoff, all of the minima fall into one group which is represented by a single node. As we increase the rate cutoff, transitions get eliminated. At some point, a critical transition gets eliminated which disconnects the minima into two groups. This is represented by the node splitting into two at the rate cutoff value. As the rate cutoff is increased further, more and more transitions are eliminated and the graph continues to bifurcate as the groups of minima further subdivide. At the base of the graph, no transitions remain, and each minimum falls into its own group. The minima can be identified at the base of the graph.

The rate disconnectivity graph for tetra-alanine shown in Figure 38 covers 23 orders of magnitude in transition rates, and contains 62357 minima.[3] Starting at the top, we see that a relatively small number of minima break away as the rate cutoff is increased to around 10 Hz. Between 10 Hz and 100 Hz, a number of large groups of minima (several thousand minima each) break away from the main branch, indicating a great deal of interesting dynamics occurring on a time scale of about 0.1 sec. Between 10^2 Hz and 10^{10} Hz relatively little happens. There seems to be two well-separated time scales with characteristic times roughly 0.1 sec and 10^{-10} sec.

The highlighted section of the rate disconnectivity graph contains a total of 3713 minima, including the extended conformation and the alpha-helical ground state. If we apply a transition rate cutoff anywhere between 10^2 Hz and 10^{10} Hz, we would find that all of the minima in the highlighted region would be connected to one another, and disconnected from all of the rest. In other words, it would take about 10^{-10} sec to make transitions between two minima within this group, and about 10^{-2} sec to make transitions out of this group. This is consistent with our solution of the Master equation (see Figure 36).

We looked for a distinguishing characteristic of the minima within this group. We found that all 3713 minima in this group satisfy the constraints

$$180^\circ \leq \phi_i \leq 330^\circ$$

for all four ϕ angles. Conversely, we found that all except for one minimum which satisfies these constraints on all four ϕ angles lies within this group. This leads us to the following conclusions:

1. Transitions involving large changes in ϕ (from within $[180^\circ, 330^\circ]$ to outside this range, or vice versa), tend to be very slow, requiring longer than 0.01 sec (sometimes much longer). This is no doubt a result of very high barriers separating these two regions of configuration space.

2. Transitions involving small changes in ϕ (i.e., those which stay within the range $[180^\circ, 330^\circ]$), and arbitrary changes in ψ, tend to be much faster, typically on the order of 10^{-10} sec. The folding of tetra-alanine from its extended conformation (bbbb) to the ground state (aaaa) falls into this catagory.

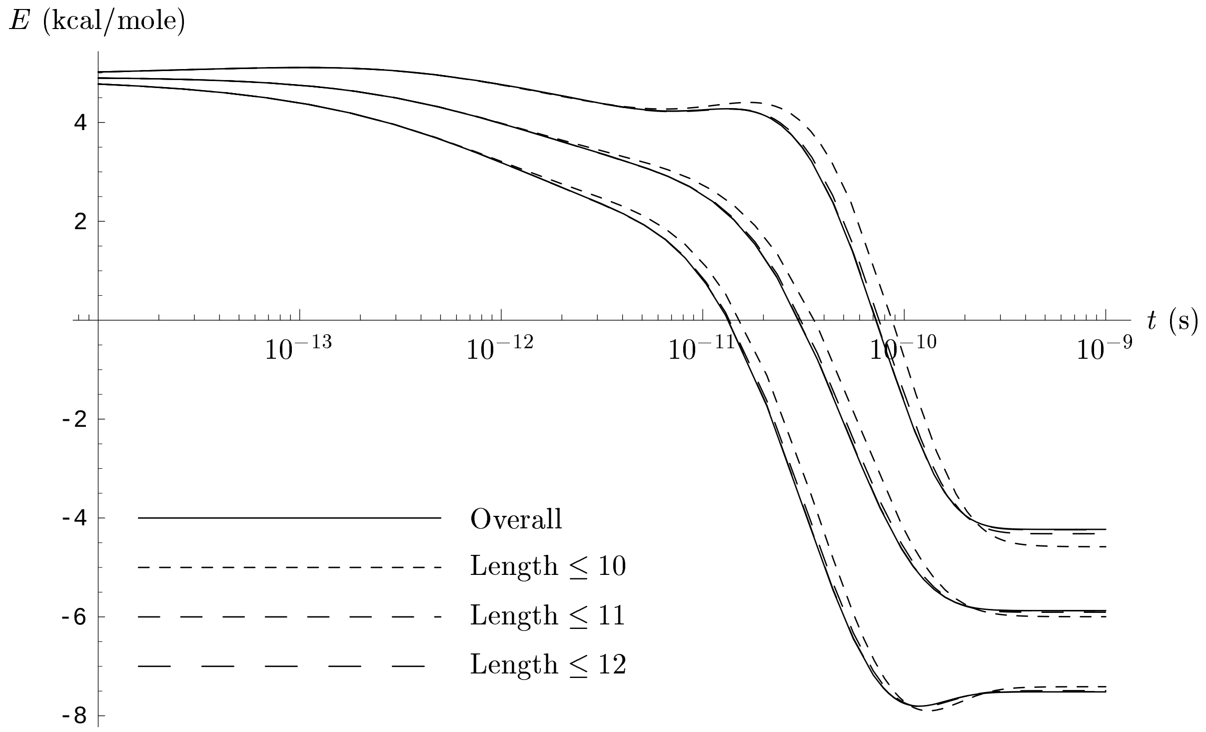

Figure 39: Time-evolution of E as a function of time (average $\pm$ one standard deviation), given that the system occupies the extended conformation at $t = 0$ sec. Various pathway length cutoffs are employed.

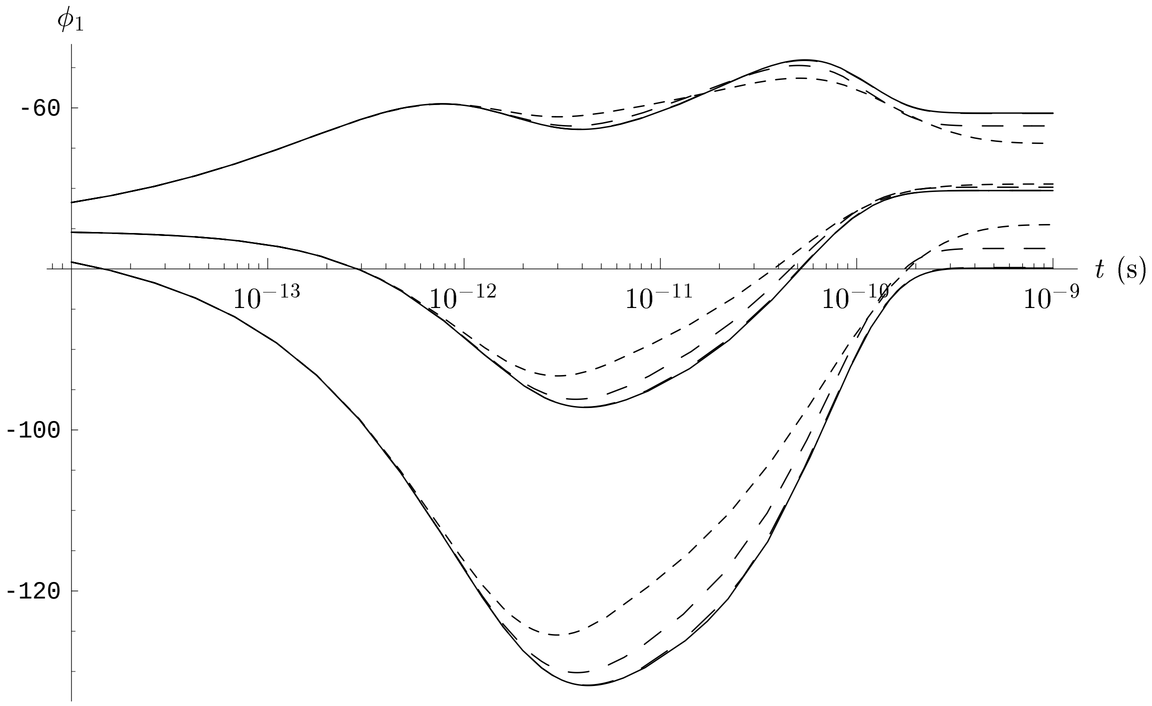

Figure 40: Time-evolution of ϕ_1 as a function of time (average $\pm$ one standard deviation), given that the system occupies the extended conformation at $t = 0$ sec. Various pathway length cutoffs are employed.

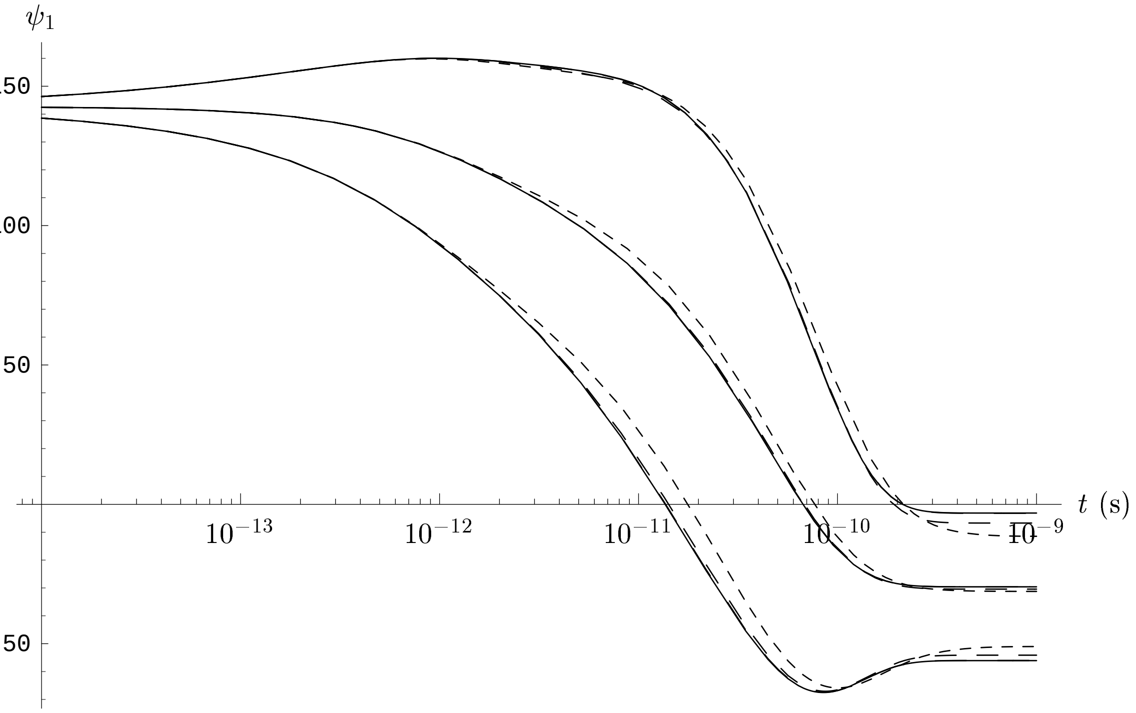

Figure 41: Time-evolution of ψ_1 as a function of time (average $\pm$ one standard deviation), given that the system occupies the extended conformation at $t = 0$ sec. Various pathway length cutoffs are employed.

5 Time Evolution of Quantities

Another way of obtaining an overall picture of the folding process of tetra-alanine is to study the time-evolution of averages of certain quantities, such as energy, dihedral angles, or distances between specific atoms. If q_j is the value of some quantity at minimum j, then $\langle q \rangle$, the average value of q, and σ_q, the standard deviation, can be calculated as a function of time with the help of the Master equation:

$$\begin{aligned} \langle q \rangle(t) &= \sum_j P_j(t) q_j = \sum_{i,j} a_i e^{\lambda^{(i)} t} u_j^{(i)} q_j \\ &= \sum_i a_i \Big(\sum_j u_j^{(i)} q_j \Big) e^{\lambda^{(i)} t} \end{aligned} \tag{96}$$

$$\langle q^2 \rangle(t) = \sum_i a_i \Big(\sum_j u_j^{(i)} q_j^2 \Big) e^{\lambda^{(i)} t} \tag{97}$$

$$\sigma_q(t) = \sqrt{\langle q^2 \rangle(t) - \langle q \rangle^2(t)} \tag{98}$$

Plots of $\langle q \rangle$ and $\langle q \rangle \pm \sigma_q$ as functions of time for $q = E, \phi_1$, and ψ_1 are given in Figures 39–41.

To obtain the correct time evolution of $\langle q \rangle$ and σ_q, it is necessary to solve the Master equation over all of the minima.[4] We can also calculate the approximate time evolution of $\langle q \rangle$ and σ_q by restricting our attention to only a certain subset of pathways. This is accomplished by restricting the minima and transition states we use to solve the Master equation to those which are visited by the selected pathways.

In Figures 39–41, we compare the overall time evolution of E, ϕ_1 and ψ_1 with the time evolution obtained by restricting our attention to pathways with various length restrictions. The deviations are rather large for a length cutoff of 10, but are much smaller for a length cutoff of 11 or 12 (the same holds true for the other ψ_i and ϕ_i angles, not shown). It appears that applying a length cutoff of 11 will yield most of the relevant pathways.

We can also determine the effect of various transition rate cutoffs on the time evolution of E, ϕ_i, and ψ_i. In Figure 42, we compare the overall time evolution of E with that obtained by restricting our attention to pathways with a length cutoff of 11 and various transition rate cutoffs. We find significant deviation from the overall time evolution only when the transition rate cutoff is increased to 10^{11} Hz (the same holds for ϕ_i and ψ_i, not shown). It appears that the most significant pathways are those of length 11 or less which satisfy a transition rate cutoff of 10^{10} Hz. There are 92216 such pathways, and they involve only 526 minima and 1696 transition states. This is significantly less than the 62373 minima and 212938 transition states that we started with.

[3] The remaining 16 minima are not connected to the main group by any transition states.

[4] Actually, we only solve the Master equation over the 3713 minima in the highlighted region of the rate disconnectivity graph shown in Figure 38. This is necessary because solving the Master equation for all 62373 minima would require diagonalizing a 62373×62373 matrix which does not fit in computer memory. Fortunately, it is also sufficient since the other minima are unreachable during times on the order of 10^{-9} sec.

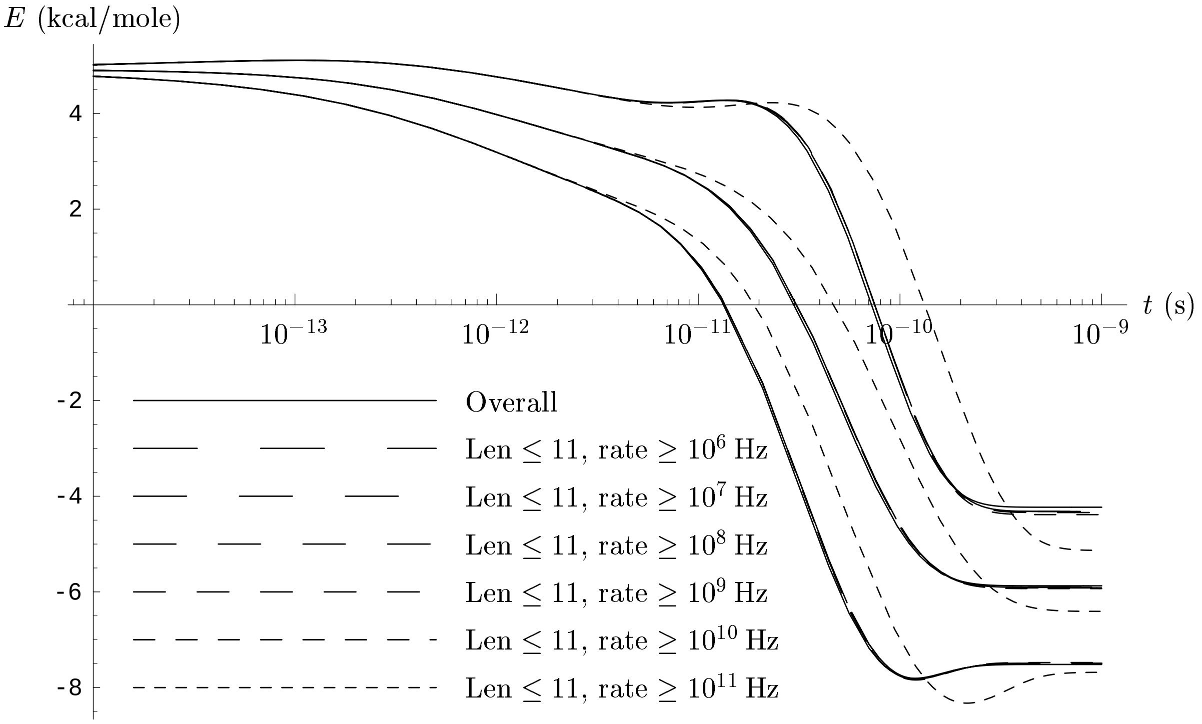

Figure 42: Time-evolution of E as a function of time (average $\pm$ one standard deviation), given that the system occupies the extended conformation at $t = 0$ sec. A pathway length limit of 11, along with various transition rate cutoffs, are employed.

6 Reaction Coordinates

It would be useful to characterize the folding process by means of determining a viable reaction coordinate. A reaction coordinate is a quantity which accurately measures the progress from the initial state to the final state. Ideally, it should be monotonic and proceed at a uniform rate along each individual pathway. If we examine the time evolution of E, ϕ_i, and ψ_i (Figures 39–41), we see that the energy and the ψ angles seem to make reasonable reaction coordinates, but the ϕ angles definitely do not. However, these plots only reveal the average progress of these quantities. What we would really like to know is which, if any, of these quantities proceeds monotonically and uniformly for *each* pathway.

To help answer this question, we developed two "reaction coordinate indicators", one which measures the monotonicity of the reaction coordinate, and the other which measures the uniformity of the reaction coordinate. For a given pathway of length N

$$\mathrm{min}_1 \to \mathrm{min}_2 \to \ldots \to \mathrm{min}_N$$

a certain quantity q takes on values

$$q_1 \to q_2 \to \ldots \to q_N$$

The two reaction coordinate indicators are d/D and D^2/S, where

$$\begin{aligned} d &= \Big|\sum_{i=1}^{N-1}(q_{i+1}-q_i)\Big| && \text{(displacement)} \\ D &= \sum_{i=1}^{N-1}|q_{i+1}-q_i| && \text{(distance)} \\ S &= (N-1)\sum_{i=1}^{N-1}|q_{i+1}-q_i|^2 && \text{(squared distance)} \end{aligned}$$

d/D measures the monotonicity of q along the given pathway, and D^2/S measures the uniformity of q along the given pathway. Both indicators take the value 1 in the ideal case.

For each of the quantities E, ϕ_i and ψ_i, we tabulated the average value and standard deviation of these two reaction coordinate indicators over the 92216 relevant pathways in Table 33. As expected, the ϕ angles perform poorly on the monotonicity test (d/D is very small), whereas the energy and the ψ angles perform reasonably well on the monotonicity test. However, none of the quantities do very well on the uniformity test: the average value of D^2/S is around 0.30 for each of the dihedral angles and around 0.48 for the energy. This suggests that changes in a given dihedral angle tend to occur in a small number of big steps, rather than in a large number of small steps. This is consistent with our earlier pathway analysis, where we found that the ψ angles tend to change one or two at a time.

It is clear that progress towards the alpha-helical ground state should not be measured in terms of a single ψ angle, but should reflect the progress of *all* ψ angles. This suggests that we might look at $\sum_i \psi_i$ as a reaction coordinate. The time evolution of $\sum_i \psi_i$ is plotted in Figure 43,

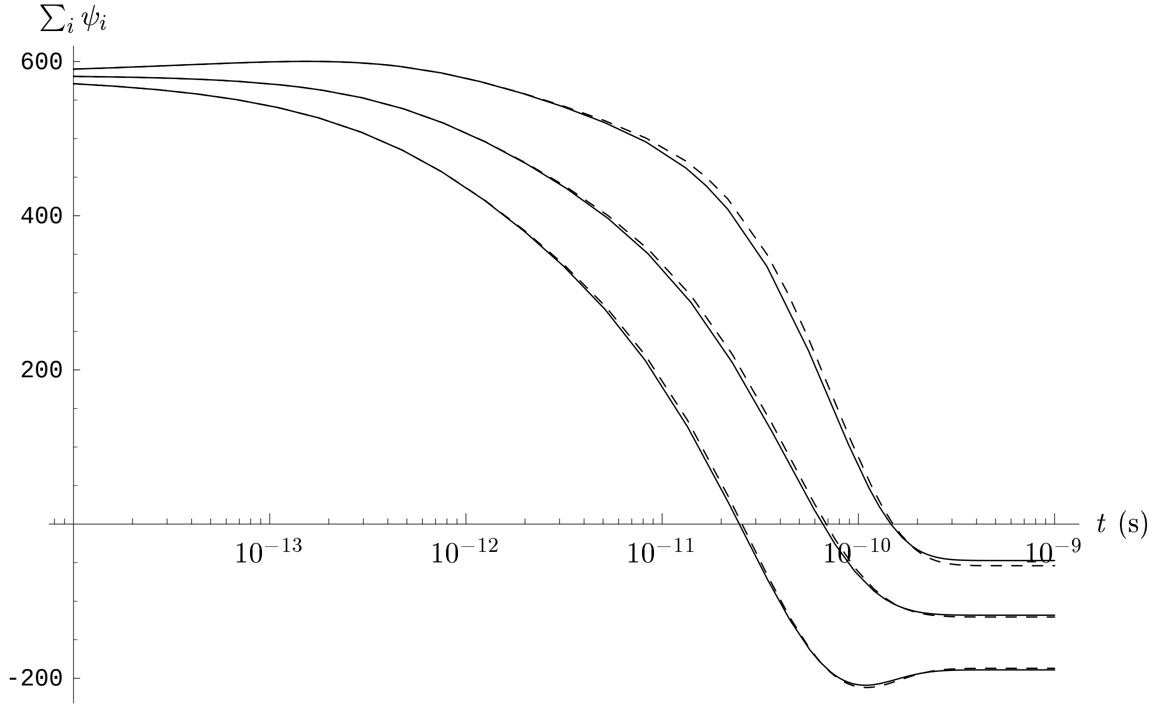

Figure 43: Time-evolution of $\sum_i \psi_i$ as a function of time (average $\pm$ one standard deviation), given that the system occupies the extended conformation at $t = 0$ sec. Solid curve shows the overall time evolution, and dotted line shows time evolution with a pathway length limit of 11 and a transition rate cutoff of 10^{10} Hz.

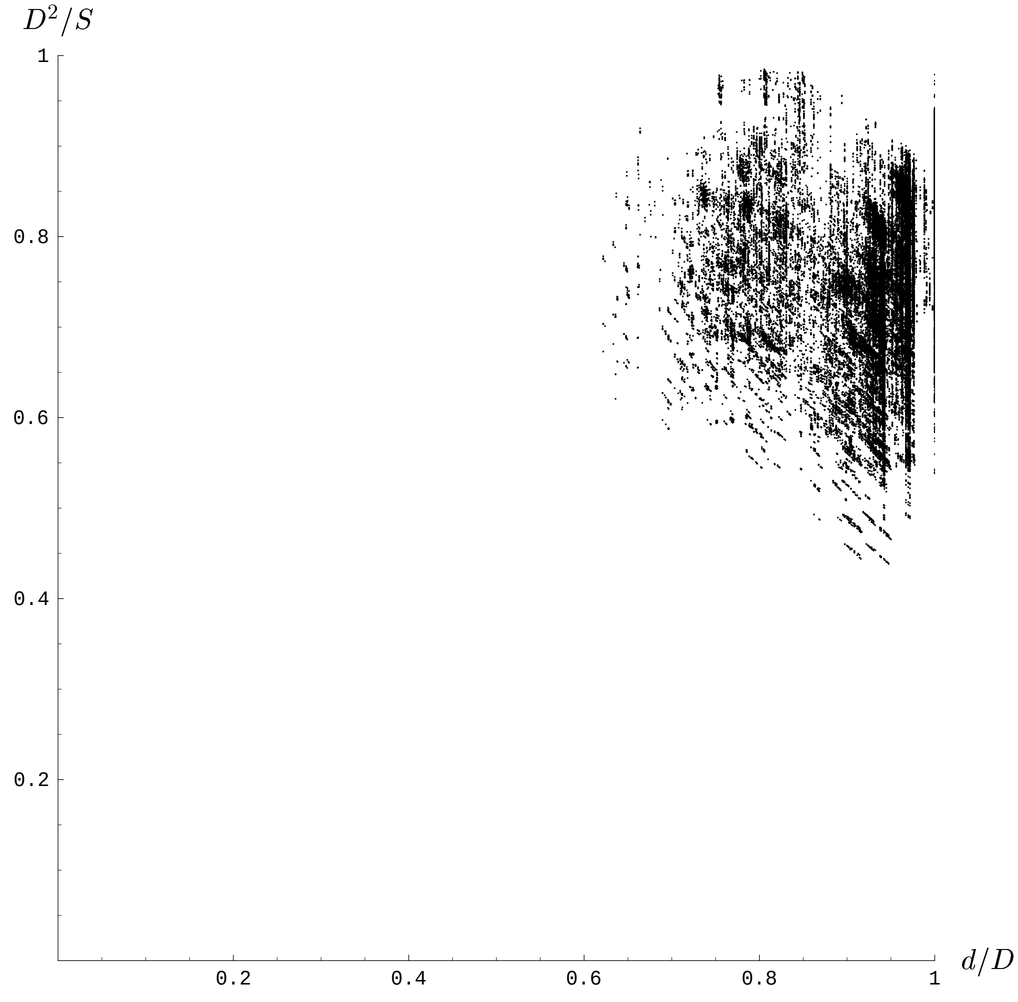

Figure 44: Scatter plot of reaction coordinate indicators for $\sum_i \psi_i$ for each pathway. Only pathways of length 11 or less with all transition rates exceeding 10^{10} Hz are used (92216 pathways).

Table 33: Average and standard deviation values of the reaction coordinate indicators d/D and D^2/S for various quantities over all pathways of length 11 or less with transition rates exceeding 10^{10} Hz from the extended conformation to the ground state of unsolvated tetra-alanine.

	d/D		D^2/S	
Quantity	avg	std	avg	std
E	0.796	0.099	0.482	0.144
ϕ_1	0.224	0.138	0.291	0.080
ψ_1	0.899	0.120	0.256	0.060
ϕ_2	0.032	0.034	0.304	0.077
ψ_2	0.850	0.100	0.283	0.051
ϕ_3	0.081	0.081	0.332	0.084
ψ_3	0.867	0.129	0.298	0.071
ϕ_4	0.046	0.038	0.302	0.075
ψ_4	0.849	0.132	0.293	0.059
$\sum_i \psi_i$	0.927	0.066	0.749	0.066
d_{α_1,α_4}	0.674	0.138	0.355	0.115
d_1	0.762	0.129	0.467	0.098
d_2	0.712	0.111	0.523	0.142
$d_1 + d_2$	0.818	0.103	0.587	0.133

and the average value and standard deviation of the reaction coordinate indicators are given in Table 33. The average value of the reaction coordinate indicators, $d/D = 0.927$ and $D^2/S = 0.749$, both indicate very strongly that $\sum_i \psi_i$ makes a good reaction coordinate. To confirm this, we constructed a scatter plot of D^2/S vs. d/D for each of the 92216 pathways, shown in Figure 44. For most of the pathways, the reaction coordinate indicators are both near 1, further suggesting that $\sum_i \psi_i$ makes a good reaction coordinate.

Further insight into the folding process may be gained by looking for a more physically significant reaction coordinate. An alpha-helix is stabilized by the formation of hydrogen bonds between the i and $i+3$ residues. Since these residues tend to be farthest apart in the extended conformation, and must be brought close together to form the hydrogen bond, it makes sense to use the hydrogen bonding distance as a reaction coordinate.

We first tried d_{α_1,α_4}, the distance between the first and fourth α-carbons. This distance is indicated in Figure 45. This distance varies from 9.079 Å in the extended conformation to 4.998 Å in the ground state. The alpha-helical ground state is not the only conformation with $d_{\alpha_1,\alpha_4} < 5.0$ Å. Of the 526 minima involved in the 92216 relevant pathways, 26 of them satisfy this inequality.

The distance between α-carbons is only a crude measure of hydrogen bonding. A more direct measure is the distance between the nitrogen-bonded hydrogen atom and the oxygen atom which shares it. It turns out there are two candidate hydrogen bonding distances, as indicated in Figure 45. These distances in the ground state are $d_1 = 1.934$ Å and $d_2 = 1.921$ Å. It turns out that neither distance alone uniquely determines the ground state. Of the 526 relevant minima, 9 of them satisfy $d_1 < 2$ Å and 7 of them satisfy $d_2 < 2$ Å. However, only the ground state satisfies both inequalities.

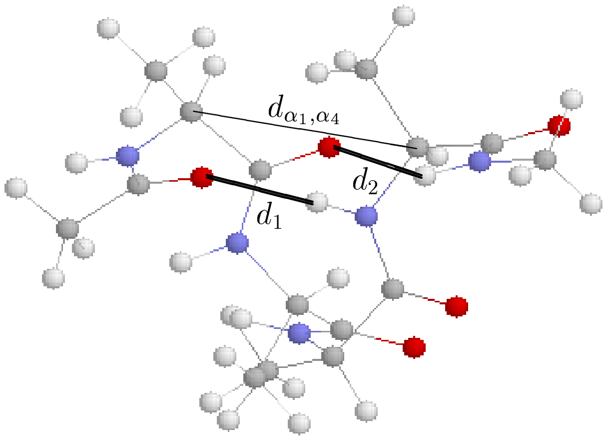

Figure 45: Alpha-helical ground state of unsolvated tetra-alanine, with the hydrogen bonds indicated.

Table 34: Eigenmode III results for solvated tetra-alanine.

	6^8 grid
Local minima	66228
1st-order saddles	195639

Apparently there are two hydrogen bonds which stabilize the alpha-helix in tetra-alanine.

We tabulated the average value and standard deviation of the reaction coordinate indicators for d_{α_1,α_4}, d_1, d_2, and $d_1 + d_2$ in Table 33. The motivation of including $d_1 + d_2$ among the distance parameters is similar to that of including $\sum_i \psi_i$. Since there are two hydrogen bonds to form, it makes sense that reaction progress should be measured by *both* hydrogen bond distances. Any of the four distance parameters would make a reasonable reaction coordinate, but $d_1 + d_2$ is clearly the best with $d/D = 0.818$ and $D^2/S = 0.587$. A scatter plot of D^2/S vs. d/D for $d_1 + d_2$ is given in Figure 46.

7 Solvated Tetra-Alanine

We next studied tetra-alanine in solvation. We used the ECEPP/3 potential energy surface coupled with the volume method for calculating solvation energies using the Reduced Radius Independent Gaussian Sphere (RRIGS) approximation.

We determined the minima and first-order saddles by applying a brute force eigenmode-following search (Eigenmode III) with a 6^8 grid of start points, just as we did for unsolvated tetra-alanine. The results of this search can be found in Table 34.

Of the 66228 minima, we found one alpha-helical conformation, min.1 (aaaa), and one extended conformation, min.874 (bbbb). The potential energy (which includes the solvation energy) and free energy (which includes contributions from the vibrational entropy) of these two states can be found in Table 35.

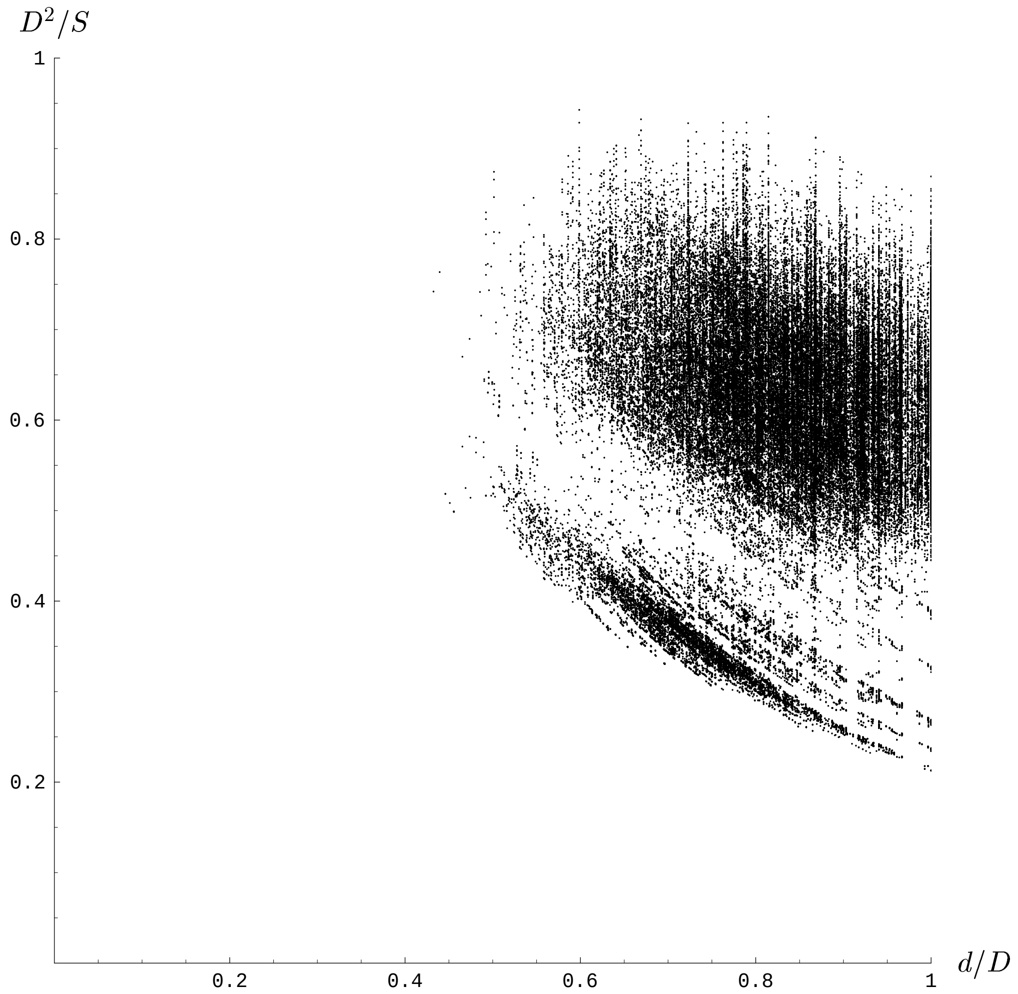

Figure 46: Scatter plot of reaction coordinate indicators for d_1+d_2 for each pathway. Only pathways of length 11 or less with all transition rates exceeding 10^{10} Hz are used (92216 pathways).

Table 35: Ground state and extended conformation of solvated tetra-alanine.

Minimum	Classification	E (kcal/mole)	F (kcal/mole)
min.1	aaaa	-35.249	-40.741
min.874	bbbb	-30.823	-41.194

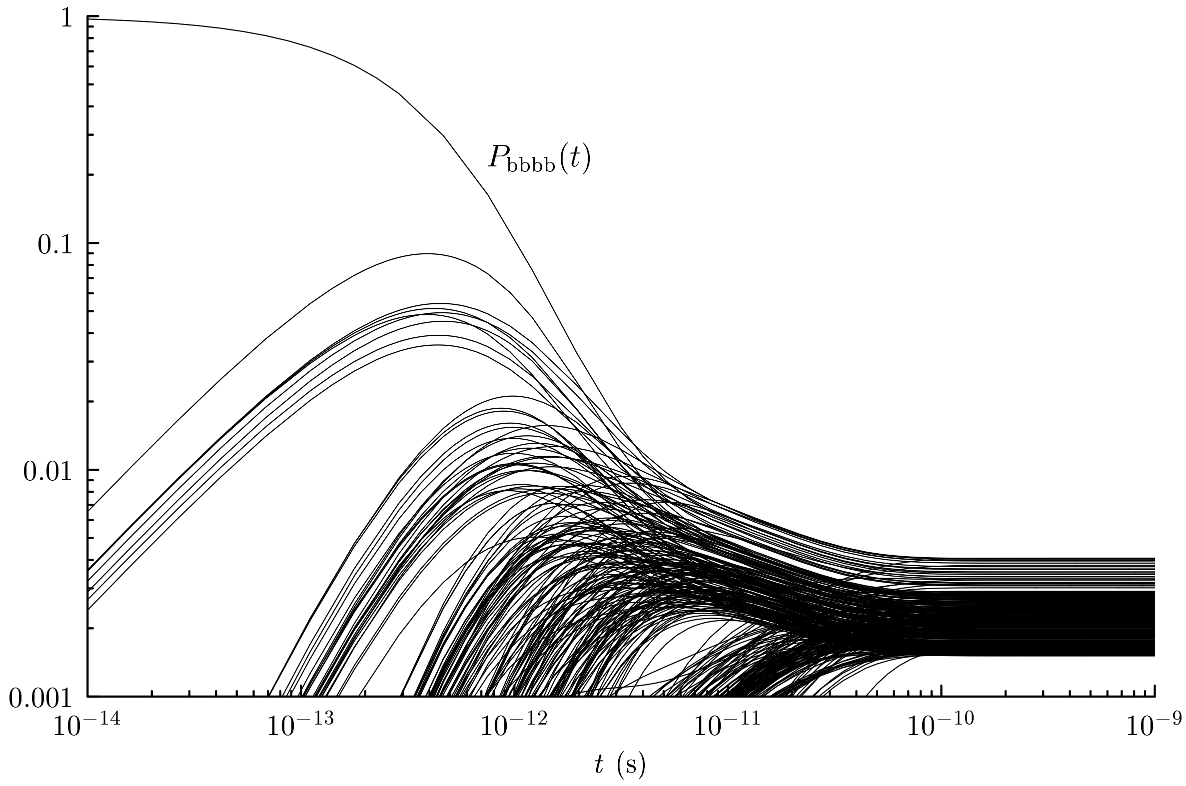

Figure 47: Time evolution of the extended conformation and the 300 lowest free energy states of solvated tetra-alanine at $T = 300\,\mathrm{K}$. No single state has an equilibrium probability which exceeds 0.004.

The first thing to notice is that, although the alpha-helical conformation has the lowest potential energy (and hence the lowest free energy at $T = 0\,\mathrm{K}$), the extended conformation has a lower free energy at room temperature ($T = 300\,\mathrm{K}$) than the ground state. The result of adding solvation energy reduces the energy gap from 11.6 kcal/mole to 4.4 kcal/mole. The entropic term in the free energy is more than enough to overpower this energy gap and reduce the free energy of the extended conformation below that of the alpha-helical ground state. This has significant implications.

We calculated the free energies of all the minima in order to determine the equilibrium probability distribution (see Section IV.C.2). We found that the several hundred lowest free energy minima have about the same free energy, and that no single minimum has an equilibrium occupation probability which exceeds 0.004. This is in stark contrast with unsolvated tetra-alanine, where the ground state had an equilibrium occupation probability of 0.748, and the lowest 3 potential energy states accounted for 0.936 of the total equilibrium probability.

As a check, we calculated the transition rate matrix for solvated tetra-alanine at $T = 300\,\mathrm{K}$, and solved the Master equation starting with the extended conformation at $t = 0$ sec. We plotted the time evolution of the occupation probabilities of the 300 lowest free energy states. That plot is given in Figure 47. The equilibrium probability distribution is achieved in about 10^{-10} sec.

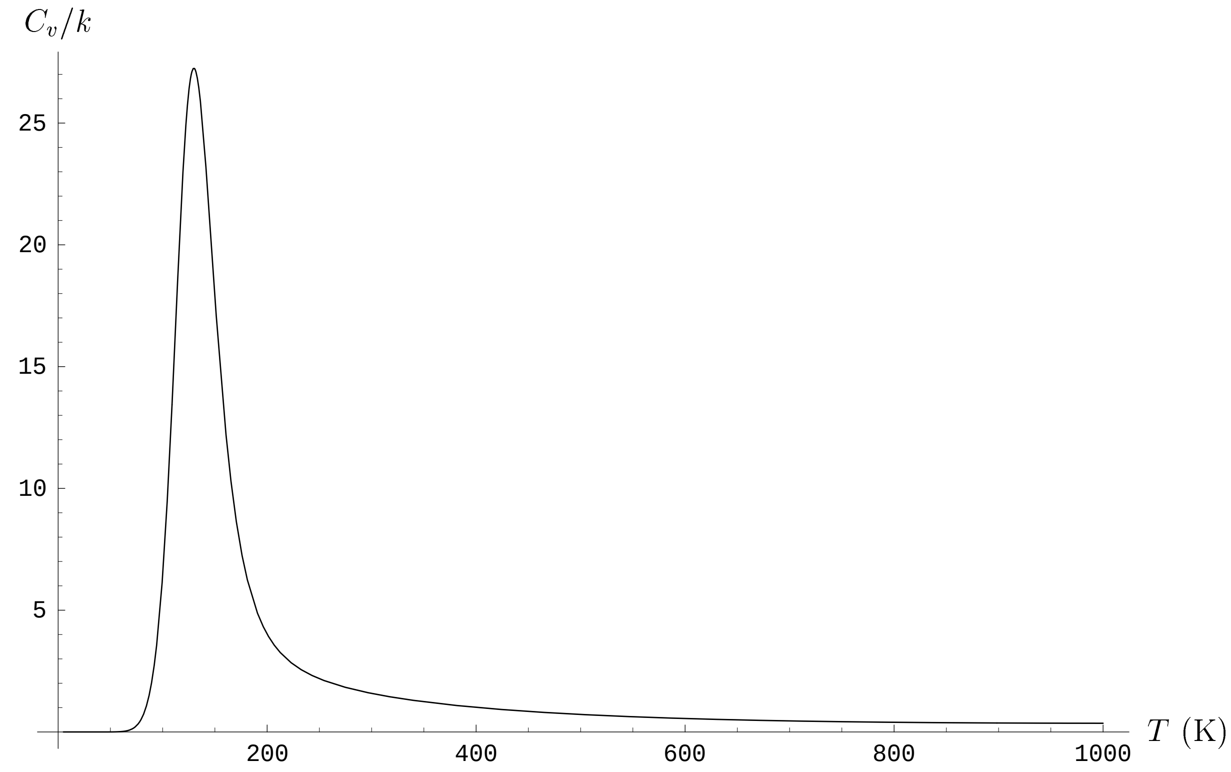

Figure 48: Heat capacity as a function of temperature for solvated tetra-alanine.

It seems likely that solvated tetra-alanine exhibits liquid-like behavior at $T = 300\,\mathrm{K}$. To be sure, we need to verify that the several hundred minima which share the equilibrium probability distribution do not occupy the same region of configuration space. If that were the case, the potential energy surface would have one deep basin with a rough bottom. The true characteristics of a liquid-like molecule is that it randomly (and quickly) samples widely distinct configurations. By plotting the distribution of minima on four (ϕ, ψ) plots (not shown), we reached the conclusion that the minima which share the equilibrium probability distribution do occupy distinct regions of configuration space.

If solvated tetra-alanine is to be liquid-like at $T = 300\,\mathrm{K}$, then there must be a phase transition. This should show up as a peak in the heat capacity versus temperature plot. The heat capacity can be calculated by calculating energy fluctuations at equilibrium

$$C_v = \frac{d}{dT}\langle E\rangle_{\mathrm{eq}} = \frac{\langle E^2\rangle_{\mathrm{eq}} - \langle E\rangle_{\mathrm{eq}}^2}{kT^2}$$

where equilibrium averages may be calculated from free energies

$$\langle q\rangle_{\mathrm{eq}} = \frac{\sum_i q_i e^{-F_i/kT}}{\sum_i e^{-F_i/kT}}$$

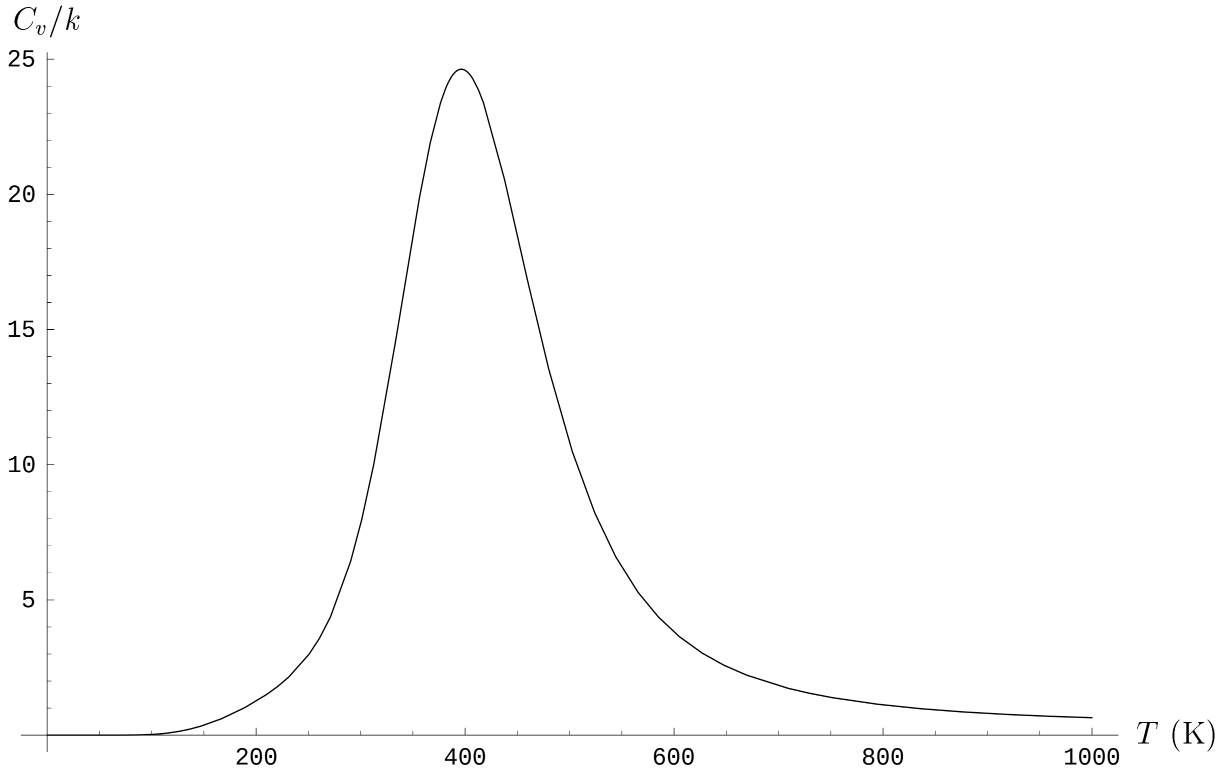

Figure 49: Heat capacity as a function of temperature for unsolvated tetra-alanine.

We calculated C_v as a function of T for temperatures ranging from (just above) 0 K to 1000 K for both solvated and unsolvated tetra-alanine. The plots are given in Figures 48 and 49. The transition temperatures are given by

$$T^{\mathrm{solv}}_{\mathrm{sol-liq}} = 130\,\mathrm{K} \qquad T^{\mathrm{unsolv}}_{\mathrm{sol-liq}} = 395\,\mathrm{K}$$

The lower transition temperature for solvated tetra-alanine can be traced back to the reduction in the energy gap between the alpha-helical ground state conformation and the other higher energy states, including the extended conformation, and indeed do explain the appearance of liquid-like behavior for solvated tetra-alanine (but not for unsolvated tetra-alanine) at $T = 300\,\mathrm{K}$.

D Overall Framework and Implementation

In this section we present the methods involved in the dynamical study of a particular peptide sequence, and discuss the implementation details of those methods. The overall framework is summarized in Figure 50. The dynamical study of a particular potential energy surface divides into two major parts: (1) the search for stationary points (minima and first and higher-order transition states), and (2) the dynamics analysis.

The stationary point search generally proceeds as follows. First, an initial sample of minima and/or transition states is generated using one of the global optimization methods (αBB, CSA, or grid-search). Additional stationary points can be generated, if needed, by performing uphill searches from minima to saddle points, or downhill searches from saddle points to minima, or by interpolating widely separated minima to locate new minima in between ("fill"). Similar minima and transition states may be combined by clustering, if desired.

Once an adequate sample of minima and transition states has been found, we begin the dynamical analysis. Connectivity between minima and transition states has already been determined by the triples calculation (i.e., downhill searches). The free energy of each stationary point is calculated (using the vibrational frequencies), and from that, the transition rates may be calculated. Then we can construct a C_v vs. T plot, determine equilibrium probability distributions, solve the Master equation, construct the rate disconnectivity graph, and perform a full pathway analysis.

1 Local Stationary Point Search Methods

Eigenmode-following search Eigenmode-following search algorithms are essentially sophisticated variations of the Newton-Raphson method applied to the equations $dV/dx_i = 0$. We employ the version introduced by Tsai and Jordan [228]. At each iteration, the Hessian may be updated by direct calculation, or by BFGS (minimum search only) or Powell updating (with occasional direct calculation) [208, 68, 88]. We generally used the Powell updating for the uphill searches and a full Hessian calculation for the downhill searches, so as to insure that the correct connectivity is determined.

The Newton-Raphson step is given by

$$\Delta \mathbf{x} = -H^{-1}\mathbf{g} = -\sum_i \frac{g_i}{b_i}\mathbf{e}_i$$

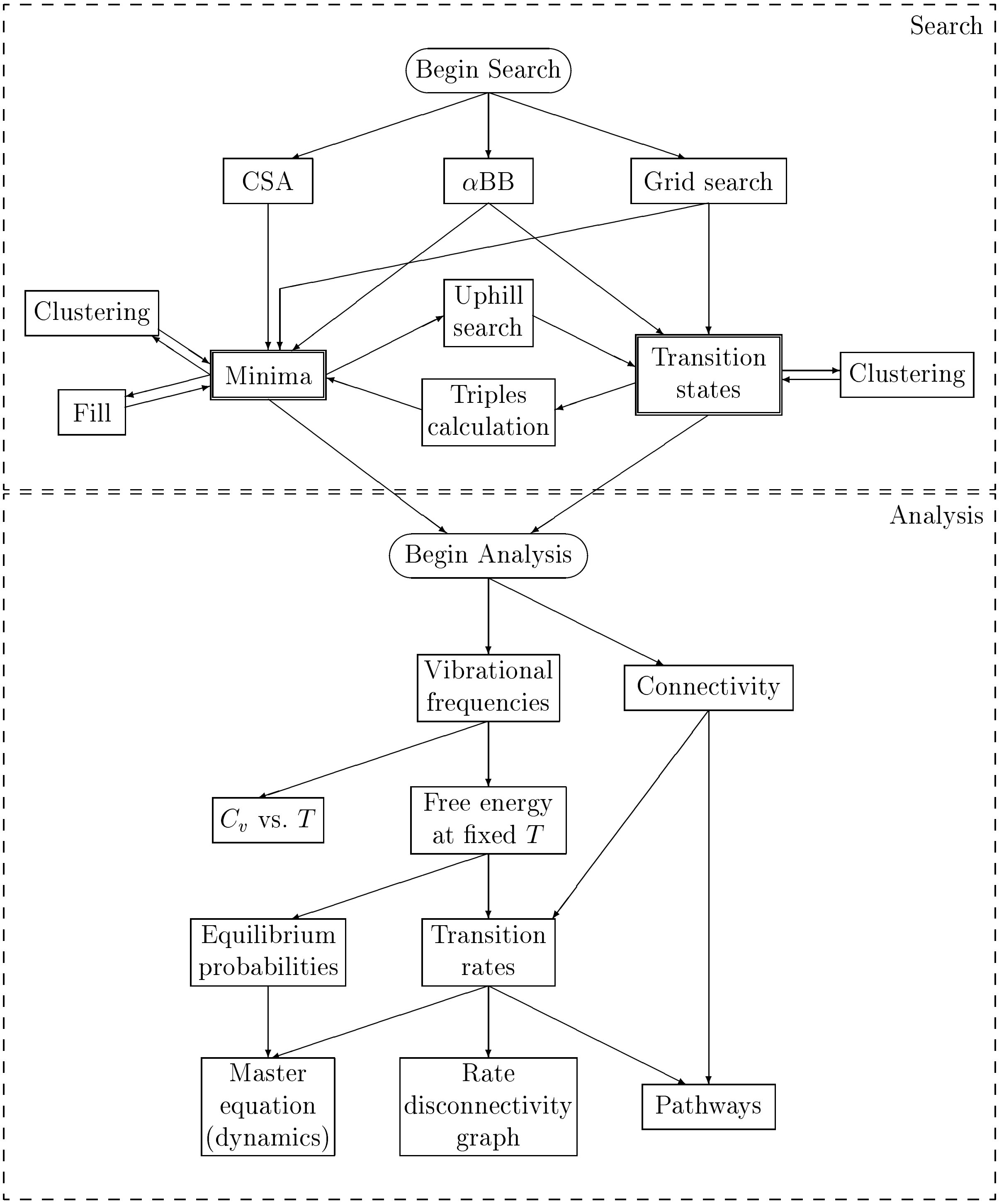

Figure 50: Overall framework for the dynamical study of a given peptide sequence.

where $\mathbf{g}$ and H are the gradient and Hessian, respectively, b_i and $\mathbf{e}_i$ are eigenvalues and eigenvectors of H, respectively, and g_i is the component of the gradient in the direction of $\mathbf{e}_i$. The Newton-Raphson algorithm tends to locate stationary points which have the same signature (i.e., number of negative eigenvalues) as the Hessian matrix at the starting point. More specifically, potential energy tends to be minimized along modes for which $b_i > 0$, and tends to be maximized along modes for which $b_i < 0$. Eigenmode-following algorithms circumvent this limitation by "shifting" some of the eigenvalues to change their sign, so that the "eigenvalues" used to construct the step have the desired signature. Thus, if a minimum is sought, then all eigenvalues are rendered positive by shifting the negative b_i's to positive values. If a first-order saddle is desired, then eigenvalues are shifted as needed so that one specific eigenvalue is negative. If the Hessian already has the required signature, no shifting takes place — the search essentially becomes a Newton-Raphson search in the vicinity of the saddle point. Eigenmode-following searches, when they converge, virtually always converge to a saddle point of the correct saddle order.

When searching for a first-order saddle from a starting point in the vicinity of a minimum, there is some question as to which eigenvalue should be shifted to a negative value (i.e., which eigenmode to follow "uphill"). There are two possible answers:

1. At each iteration, follow the mode with the smallest eigenvalue.
2. Choose a specific mode at the starting point, and continue to follow that "same" mode as each step is taken. Eigenmodes at each subsequent step are identified with eigenmodes at previous steps by maximum overlap.

In the first case, it is automatically the case that at any point where the Hessian has the correct signature, no eigenvalue shifting takes place at all, and the Newton-Raphson step is taken. In the second case, a specific mode is selected, which often does not start out as the lowest eigenvalue mode. However, as the selected mode is driven uphill, its eigenvalue decreases, eventually causing that mode to overtake the other modes in becoming the lowest eigenvalue mode. Eventually the eigenvalue is driven to a negative value, after which the first-order saddle will be found.

SUMSL The SUMSL algorithm [76], which is made available as part of ECEPP/3 [180], is designed to find local minima in the vicinity of a starting point. It employs the BFGS updating method for the Hessian. It is specifically designed for minimum searches, and as such, is generally much more efficient than the eigenmode-following algorithm.

2 Methods for Finding Minima and First-Order and Higher-Order Transition States

αBB Stationary points of all orders are generated by solving the stationary conditions

$$\frac{dV}{dx_i} = 0 \qquad i = 1, \ldots, N_x$$

using the αBB method described in Sections II.B and IV.B. This algorithm offers a theoretical guarantee of enclosing all solutions within the starting region in a finite amount of time.

CSA The Conformational Space Annealing (CSA) algorithm attempts to reach the global minimum (free-)energy conformation by a combination of genetic, annealing, and build-up algorithms [144, 145, 143]. The user provides an initial bank of minima (usually by locally minimizing randomly selected points). Seed points are selected from the bank and modified according to prespecified rules. The modified points are then minimized by local search, and then considered for introduction into the bank, possibly replacing a point which is already there. If the candidate point falls within a certain "cutoff distance" from any other point in the bank, the candidate point and the bank point closest to it are compared. Otherwise, the candidate point is considered to be in its own "class", and is compared with all other points in the bank. In either case, the highest (free-)energy point is discarded. The "cutoff distance" is initialized to $1/2$ the average distance between points in the initial bank, and is annealed down by a fixed factor every iteration.

Termination conditions include one or more of the following:

1. Iteration count limit.

2. Round limit. Each round ends after every point in the current bank has been used as a seed point.

3. (Free-)energy lower limit.

4. Update counter limit. The "update counter" is incremented whenever the fraction of candidate points which actually make it into the bank is sufficiently small for a given number of minimizations.

5. Stop file. The user can stop the algorithm by creating a special file whose existence is checked each iteration.

Virtually all of the effort is spent performing the local minimizations of the modified seed points. The parallel version of this algorithm divides the modified seed points among all of the processes (including the master process) to be minimized in parallel. The master process handles the rest of the algorithm.

Grid search A sampling of stationary points of a specified order (first-order saddle or minima, generally) is found by initiating an eigenmode-following search from each point on a specified grid to a saddle point of the specified order. After the searches are completed, duplicate points are thrown out. This algorithm requires one search for each grid-point, and thus the time requirements depend exponentially on the number of variables for which alternative values are provided for the grid. It is therefore unsuited for large problems, but yields good results for small problems (e.g., tetra-alanine, discussed in Section IV.C).

The parallel version of this algorithm divides the grid-points among all of the processes (including the master process), which then perform the searches. The results are sent back to the master process.

Uphill search First-order saddle points are found by performing eigenmode-following searches "uphill" from each minimum. For every minimum and every choice of eigenmode, an initial step is taken along that mode (in each of the two possible directions), followed by an eigenmode-following search along that mode to a first-order saddle. A total of $2N$ searches are required for each minimum, where N is the number of eigenmodes. One may alternatively restrict the number of modes followed for each minimum. The resulting saddles are collected and duplicates are removed.

The parallel version of this algorithm divides the minima among all of the processes (including the master process), which then perform all of the required searches. The results are sent back to the master process.

Triples calculation The connection between first-order saddles and the minima they connect are found by performing a minimization on each side of the saddle point. An initial step from the saddle point is taken in each of the two directions along the eigenmode corresponding to the negative eigenvalue, each followed by a minimization. Minima found this way are compared with minima that have been found previously, and duplicates are discarded in favor of the previously found minima. This algorithm may also be used to locate previously unknown minima by downhill search from a saddle point, in which case the new minima are retained.

The parallel version of this algorithm divides the saddle points among all of the processes, which then perform the necessary minimizations. As the saddles and minima are sent back to the master process, duplicate minima are discarded in favor of previously determined minima.

Fill "Fill" refers to the act of filling in a "scaffolding" of minima (such as might be obtained by a CSA run) by searching for additional minima between pairs of minima found in the initial set. The reason why this may be necessary is because the minima generated by the CSA algorithm are often too far apart for connections between them to develop after a single uphill/downhill search (this is practically by design of the CSA algorithm, which spreads itself thin so as to sample a large portion of the conformational space).

A "distance cutoff" and a "coordination number" are provided, as well as an initial set of minima. Ideally, this algorithm will first cluster the points according to the distance cutoff (i.e., split the points into equivalence classes, where two points are equivalent if they can be connected to one another by a "path" involving points which are within the cutoff distance from each other). Then each cluster will be paired with the N clusters closest to it in distance, where N is the coordination number. For every pair of clusters generated this way, additional points will be added along the line joining the two clusters (more specifically, along the line joining the two representative points in the two clusters which are the least distance apart). The points will be uniformly spaced and the number of points chosen is the least number which results in each point being within the cutoff distance from its nearest neighbor. The new points can then be used as starting points for a minimization.

For practical reasons, the algorithm actually proceeds as follows. Every point is considered in turn. Distances from that point to every other point are first determined, and then sorted. Connections between this point and all points which are within the cutoff distance are noted, so that equivalence classes may be determined. Then the pairs consisting of this point and each of the next N points (N is the coordination number), along with their mutual distances is added to a "pairs" list. After all points have been considered, and the equivalence classes are determined, (class-wise) duplicate pairs are discarded from the "pairs" list, and then the points are generated as described above.

The most CPU-intensive part of the algorithm is the generation of a list of distances between a given point and every other point, and the sorting of that list. In the parallel version of this algorithm, the master process sends the set of points to each slave process so that they will know what to do. While the master process is carrying out the remainder of the algorithm, each slave process calculates the distances between a given point and every other point, and then sorts the list. As each point is considered in turn, the master process cycles through the slave processes, receiving the needed distances from each one.

Clustering The number of minima and transition states can be reduced by "clustering" them, that is, by identifying points which lie within a specified distance of one another with a single point. The first point in the set of points to be clustered is selected as a cluster center and compared with every other point. Points within a certain cutoff distance from the selected cluster center are identified as belonging to that cluster and taken out of circulation. The next point in the set that is not yet part of any cluster is selected as the next cluster center, and is compared with all other points not yet part of any cluster. The algorithm continues this way until all points have been assigned to a cluster.

Note that the clusters generated by this algorithm have the property that the cluster centers used to generate them appear earlier than all of the other points in the cluster. Thus, by first sorting the set in increasing order of potential energy, we can guarantee that each cluster will be represented by its lowest energy member, and in particular, that the global minimum energy point will be among the cluster centers.

Minima should be clustered first using the algorithm as described above. The connectivities between the transition states and the minima they connect should then be redefined so that transition states connect the cluster centers associated with the minima they actually connect. Then the transition states can be clustered using the algorithm as described above with one additional caveat: one transition state cannot be identified as belonging to a cluster centered by another transition state unless they connect the same two minima (clusters).

The most CPU intensive part of the algorithm is the calculation of distances between selected cluster centers and all other non-clustered points. The parallel version of this algorithm runs as follows. The points are first sorted, and then shipped from the master process to each slave process. The master process sends the first cluster center to one of the slave processes which begins comparing that point to all of the remaining points in the set. As cluster matches occur, results are reported back to the master process which records the clustering information. The master process continues sending cluster centers to available slave processes and awaiting reports of clustering until all points have been clustered.

The situation is complicated by the fact that the master process cannot send a new cluster center to another slave process until it is established that the potential cluster center does not belong to a cluster defined by a previous cluster center. As long as each slave process performs the comparisons in order, the master process will be able to deduce that the next unclustered point should be regarded as a new cluster center as soon as all active slave processes have reported progress beyond that point. To facilitate this process, each slave process will report its progress back to the master process at well-defined intervals,[5] in addition to those instances where a cluster match is found.

3 Methods for Analyzing the Potential Energy Surface

Vibrational frequencies calculation The vibrational frequencies are determined by solving the generalized eigenvalue problem

$$(H - (2\pi\nu)^2 I)x = 0$$

where H is the Hessian and I is the generalized inertia tensor, defined so that the kinetic energy of the system is given by

$$K = \frac{1}{2}\sum_{i,j} \frac{dx_i}{dt} I_{ij} \frac{dx_j}{dt}$$

The inertia tensor is calculated by first calculating dr_j/dx_i for $j = 1, \ldots, 3N_a$ and $i = 1, \ldots, N_x$ by finite differencing, and then using the following formula:

$$I_{ii'} = \sum_{j=1}^{3N_a} m_j \frac{dr_j}{dx_i} \frac{dr_j}{dx_{i'}}$$

where m_j is the mass of the $j/3$-th atom.

This makes use of the Cartesian coordinate functions $\mathbf{r}(\mathbf{x})$. The formulae above depend on the Cartesian coordinates being physically correct. Unfortunately, most methods of generating Cartesian coordinates from generalized coordinates (in our case, dihedral angles) involve fixing the positions and orientations of specific atoms, which leads to the introduction of unphysical forces and torques being applied to the molecule. We eliminate these unphysical forces by augmenting the set of generalized coordinates to include overall translation and rotation coordinates, calculating the vibrational frequencies using the above methods, and then discarding the 6 zero-mode frequencies (which must exist). The resulting vibrational frequencies are physically correct.

Vibrational frequencies can be computed at the end of an eigenmode-following search at little cost, since the Hessian has already been generated. Alternatively, the vibrational frequencies can be calculated all at once after the minima and saddles have all been found. In the latter case, the calculation can be run in parallel by distributing the work to each process, having them calculate the frequencies, and then having them pass the results back to the master process.

[5] A geometric sequence is used so as to generate a number of early reports without generated an enormous number of total reports. Thus, reports are sent back after comparing the cluster center to the next 1, 5, 5^2, 5^3, ... points.

Free energy calculation The free energy for a given stationary point is defined as follows:

$$F = E - TS_{\text{vib}}$$

The vibrational entropy is calculated from the vibrational frequencies by employing the Classical Harmonic Oscillator approximation

$$S_{\text{vib}} = -k \ln \prod_i \frac{h\nu_i}{kT}$$

where the product is taken over all vibrational frequencies. For saddle points of order 1 or higher, the negative eigenvalue modes are not counted as "vibrational modes".

Other methods of calculating the vibrational entropy exist, but are not currently implemented. Perhaps the simplest is the Quantum Harmonic Oscillator approximation:

$$S_{\text{vib}} = -k \ln \prod_i 2 \sinh \frac{h\nu_i}{2kT}$$

Anharmonic methods exist in the literature [18, 19].

Equilibrium probabilities Equilibrium probabilities are calculated from the contribution to the partition function from each minimum, which can be expressed in terms of its free energy:

$$P_i = \frac{e^{-F_i/kT}}{\sum_j e^{-F_j/kT}}$$

The minimum free energy over the entire system is first subtracted off in order to prevent overflow/underflow problems which could arise from modest non-zero free energies (positive or negative).

Average values (as well as standard deviations) of any quantity can now be computed at equilibrium:

$$\begin{aligned} \langle q \rangle &= \sum_i q_i P_i \\ \sigma_q &= (\langle q^2 \rangle - \langle q \rangle^2)^{1/2} \end{aligned}$$

Temperature derivatives are also possible:

$$\frac{d\langle q \rangle}{dT} == \frac{\langle qE \rangle - \langle q \rangle \langle E \rangle}{kT^2}$$

assuming that q_i does not depend explicitly on temperature. In particular, the specific heat $C_v = d\langle E \rangle / dT$ can be calculated.

Transition rates Transition rates are computed by Rice-Ramsperger-Kassel-Marcus (RRKM) theory. Each transition state is associated with two rates:

$$\begin{aligned} W_{i \to \text{ts} \to j} &= \frac{kT}{h} e^{-(F_{\text{ts}} - F_i)/kT} \\ W_{j \to \text{ts} \to i} &= \frac{kT}{h} e^{-(F_{\text{ts}} - F_j)/kT} \end{aligned}$$

These rates are collected together in a (sparse) matrix:

$$W_{ij} = \sum_{\text{ts}} W_{j \to \text{ts} \to i}$$

Time-dependent probabilities (Master equation) The time development of occupation probabilities can be determined by solving the Master equation:

$$\frac{dP_i}{dt} = \sum_j W_{ij} P_j - (\sum_j W_{ji}) P_i = \sum_j w_{ij} P_j$$

where

$$w_{ij} = \begin{cases} W_{ij} & \text{(if } i \neq j\text{)} \\ -\sum_{i'} W_{i'i} & \text{(if i = j)} \end{cases}$$

Solving the Master equation involves determining the eigenvalues and eigenvectors of the (non-symmetric, but easily symmetrizable) matrix w:

$$\sum_j w_{ij} u_j^{(k)} = \lambda^{(k)} u_i^{(k)}$$

The occupation probabilities as a function of time can be computed (and, e.g., plotted):

$$P_j(t) = \sum_k a_k e^{\lambda^{(k)} t} u_j^{(k)}$$

where the coefficients a_k are determined from the initial conditions $P_j(0)$. The time constants are determined from the eigenvalues

$$\tau_k = -1/\lambda^{(k)}$$

One of the eigenvalues is zero, which corresponds to the equilibrium probability distribution ($\tau = \infty$). The remaining eigenvalues will be negative.

Average values (as well as standard deviations) of any quantity can now be computed as a function of time (and, e.g., plotted):

$$\begin{aligned} \langle q \rangle (t) &= \sum_j q_j P_j(t) = \sum_k a_k (\sum_j q_j u_j^{(k)}) e^{\lambda^{(k)} t} \\ \sigma_q(t) &= (\langle q^2 \rangle (t) - \langle q \rangle (t)^2)^{1/2} \end{aligned}$$

Solving the Master equation requires the diagonalization of a matrix whose size is the number of minima in the system. This is an extraordinarily expensive operation, and may be prohibitive in both space and time resources required. A 4000×4000 matrix requires 128 megabytes of storage and generally requires about a day of CPU time to diagonalize. There is no parallel algorithm available for this operation.

Pathways Each transition state connects two minima on the potential energy surface. A pathway between two minima is defined as a series of such connections:

initial state $\rightarrow$ ts $\rightarrow$ min $\rightarrow$ ts $\rightarrow \ldots \rightarrow$ ts $\rightarrow$ min $\rightarrow$ ts $\rightarrow$ final state

The set of all (non-looping) pathways from one minimum to another can be found by an exhaustive search. We begin at the initial state and move to each minimum which is connected to the initial state. For each such minimum, we recursively explore all minima connected to that minimum, taking care not to visit a given minimum more than once along the same pathway. When the final state is reached, the pathway is reported. When all possible routes have been explored, the algorithm terminates.

For any reasonably sizable system of minima and transition states, the number of possible non-looping pathways between any two minima is likely to be prohibitively large. There are several criteria that can be applied to reduce the number of pathways:

1. Monotonicity in any specified quantity (such as energy or free-energy). Transitions are ignored if they violate the proposed monotonicity.

2. Maximum length (i.e., maximum number of minima, including the initial and final state, visited along the pathway).

3. Minimum transition rate. Transitions are ignored if they are slower than this cutoff rate.

The following information is available during a pathway calculation:

1. The set of minima and/or transition states visited along the way by at least one of the pathways.

2. Transition rates for each transition taken along a given pathway.

3. An overall "transition time" for a given pathway. This is determined by solving the Master equation over the minima and transition states involved in that one pathway alone, and using the lifetime of the longest lived transient probability eigenvector.

4. Values of any number of quantities for each minimum visited along a given pathway.

5. The two reaction coordinate indicator values associated with any number of quantities along each given pathway (explained below).

6. The average value and standard deviation of any number of quantities over all pathways, at a fixed position along the pathways.

7. The average value and standard deviation of the two reaction coordinate indicators over all the pathways (explained below).

For a given pathway

$$\mathrm{min}_1 \to \mathrm{min}_2 \to \ldots \to \mathrm{min}_N$$

a certain quantity q takes on values

$$q_1 \to q_2 \to \ldots \to q_N$$

To help determine if q would make a good reaction coordinate, we developed two "reaction coordinate indicators". They are d/D (monotonicity) and D^2/S (uniformity), where

$$
\begin{aligned}
d &= \Big| \sum_{i=1}^{N-1} (q_{i+1} - q_i) \Big| \\
D &= \sum_{i=1}^{N-1} |q_{i+1} - q_i| \\
S &= (N-1) \sum_{i=1}^{N-1} (q_{i+1} - q_i)^2
\end{aligned}
$$

An ideal reaction coordinate varies both monotonically (same direction) and uniformly (in equal steps) from its initial value to its final value. Both reaction coordinate indicators take on the value of 1 in this ideal case. Values less than 1 indicate non-ideality.

Less detailed information about the connectivity of the minima is also available. The level of connection between two minima is defined as the minimum length pathway which connects them. The level of connection between a given minimum and all other minima can be generated iteratively as follows. First, start off by marking the given minimum as level 1 with all other minima marked (temporarily) as unreachable. For each level n, starting with $n = 1$, we follow each minimum marked as level n to all the minima they are connected to. For each such connected minimum, if it is yet to be marked as reachable, it is marked as level $n+1$ (if it is marked already, then a shorter pathway has already reached it). We continue on with level $n+1$, stopping whenever no additional minima are marked for a given level.

This procedure may be used to determine the connection component which contains a given minimum (i.e., the set of minima connected to the given minimum by *any* length pathway). By iteratively applying this procedure, the minima can be divided into connection components.

It should be noted that pathway traversal can be substantially optimized when a length restriction is given. First, the level of connection between the final state and all other minima is determined. Then, for every transition considered during the pathway search, it is determined whether or not the final state could possibly be reached in the proper number of steps. If it is not possible according to the precalculated level of connection, the transition is avoided.

Rate disconnectivity graph Minima can be classified into connection components. If a transition rate cutoff is applied, transition states may be eliminated if the transitions they represent occur too slowly. In this case, the number of connection components may increase. The rate-dependent connectivity information can be summarized by drawing a rate disconnectivity graph. One starts off at the top of the graph with a low rate cutoff, in which case the minima are separated into their connection components. As the rate cutoff is increased, transition states get eliminated from consideration. At some critical value of the transition rate cutoff, a critical transition state gets eliminated, causing one of the connection components to divide in two. As the rate cutoff is increased further, more and more transition states are eliminated from consideration, causing further bifurcation of connection components. At the highest rate cutoffs, no transition states remain, and all minima occupy their own connection component. Minima can be identified at the base of the graph.

The rate disconnectivity graph is built from the bottom up. Each minimum starts off by occupying the leaf node of its own tree. Connectivities between pairs of minima are sorted in decreasing order of transition rate, so that the highest transition rates will be considered first.[6] For each such pair of minima, we locate the subtrees generated so far which contain each of the two minima. If the two minima already belong to the same subtree, nothing happens. If the two minima belong to different subtrees, those two subtrees are joined by a bifurcation node, which is labeled with the transition rate. The rate disconnectivity graph will be completed after each transition has been considered, at which point there will be one tree for each connection component.

Once the rate disconnectivity graph is constructed, one can walk along the nodes in the tree, print a subtree in text format, or write Mathematica code which plots the rate disconnectivity graph in graphical form.

E Perspectives and Future Work

In this section, we discuss our ongoing efforts to elucidate the folding mechanism of beta-hairpin and beta-sheet structures by studying one of the short peptides which has been recently discovered to form such structures in the native state.

Our first task centered on the selection of an appropriate peptide sequence and a potential energy surface. Our initial efforts were focussed on a 12-residue designed sequence using the ECEPP/3 potential energy surface with an additional solvation term using the volume method. Unfortunately, we were unable to locate a low-energy hairpin structure, and upon further investigation, discovered that the lowest energy state of this system was an alpha-helix. It seems that ECEPP/3 is unable to predict the beta-hairpin structure of this peptide sequence. So we checked other peptide sequences as well as other potential energy surfaces to see if we could predict a beta-hairpin fold. We eventually found success with the second beta-hairpin segment of Protein G (residues 41–56) using the Effective Energy Function (EEF1) [140], which is the CHARMM potential plus a solvation term based on a Gaussian solvent exclusion model.

Segment 41–56 of Protein G is a 16-residue peptide which has been determined experimentally to fold into a beta-hairpin in aqueous solution [26]. A schematic of this hairpin structure is depicted in Figure 51. The hairpin structure is stabilized by the formation of three pairs of hydrogen bonds as indicated in Figure 51. The corresponding distances between the C_α atoms are designated d_1, d_2, and d_3, and will play an important role in our analysis of this molecule.

The potential energy surface we employ for this peptide in aqueous solution is split into two terms:

$$E = E_{\mathrm{pep}} + E_{\mathrm{solv}}$$

where E_{pep} includes the peptide intra-molecular interactions, and E_{solv} includes the peptide-solvent and solvent-solvent interactions.

[6] The transition rate associated with a given pair of connected minima is by default the maximum of the two transition rates associated with that connection. The minimum transition rate can be selected instead.

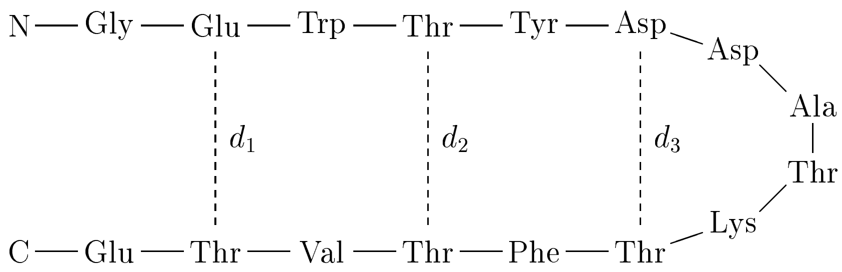

Figure 51: A schematic of Protein G (41–56) in its hairpin conformation. The dotted lines indicate hydrogen bonds, and the distances d_1, d_2, and d_3 refer to the distances between the C_α atoms.

The intra-molecular interaction term, E_{pep}, is modeled with the CHARMM22 potential energy function, an all-atom potential which takes the general form [114, 31]

$$\begin{aligned} E_{\text{pep}} \;=\; & \sum_{\text{bonds}} K_b(b-b_0)^2 + \sum_{\text{Urey-Bradley}} K_{UB}(S-S_0)^2 + \sum_{\text{angles}} K_\theta(\theta-\theta_0)^2 + \\ & \sum_{\text{dihedrals}} K_\phi(1+\cos(n\phi-\delta)) + \sum_{\substack{\text{improper}\\ \text{dihedrals}}} K_\omega(\omega-\omega_0)^2 + \\ & \sum_{\text{non-bond}} \left\{\epsilon_{ij}\left[(R_{ij}^{\min}/r_{ij})^{12} - 2(R_{ij}^{\min}/r_{ij})^6\right] + (q_i q_j)/r_{ij}\right\} \end{aligned} \tag{99}$$

The quantities b, S, θ, ϕ, ω are the bond length, Urey-Bradley distance, bond angle, dihedral angle, and improper dihedral angle, respectively, with the zero subscript representing equilibrium values. The parameters have been determined empirically and are given in [114].

The solvation term, E_{solv}, is based on the Gaussian solvent exclusion model, which takes the general form [140]

$$\begin{aligned} E_{\text{solv}} &= \sum_i \Delta G_i^{\text{solv}} \\ &= \sum_i \left\{\Delta G_i^{\text{ref}} - \sum_{j\neq i} \frac{\alpha_i e^{-((r_{ij}-R_i)/\lambda_i)^2}}{4\pi r_{ij}^2} V_i\right\} \end{aligned} \tag{100}$$

where r_{ij} is the distance between atoms i and j. The parameters ΔG^{ref}, α_i, R_i, λ_i and V_i can be found in [140]. In addition, partial charges for several atoms in charged residues have been modified, effectively neutralizing the sidechains in the CHARMM22 potential.

To simplify calculations, we fix the bond lengths and bond angles to their equilibrium values according to the CHARMM22 parameters, allowing only the ϕ, ψ, ω, and χ dihedral angles to vary. This reduces the number of degrees of freedom from $3N_a - 6 = 735$ to $N_h = 88$. Energy values, as well as the Cartesian gradient and Hessian matrix were computed by the TINKER software package [192]. The Cartesian gradients and Hessians were converted to torsional gradients and Hessians by methods developed in our computer lab. All in all, one Hessian evaluation requires approximately 0.50 sec of CPU time on a 600 MHz pentium machine running linux, where the bulk of the calculations were performed.

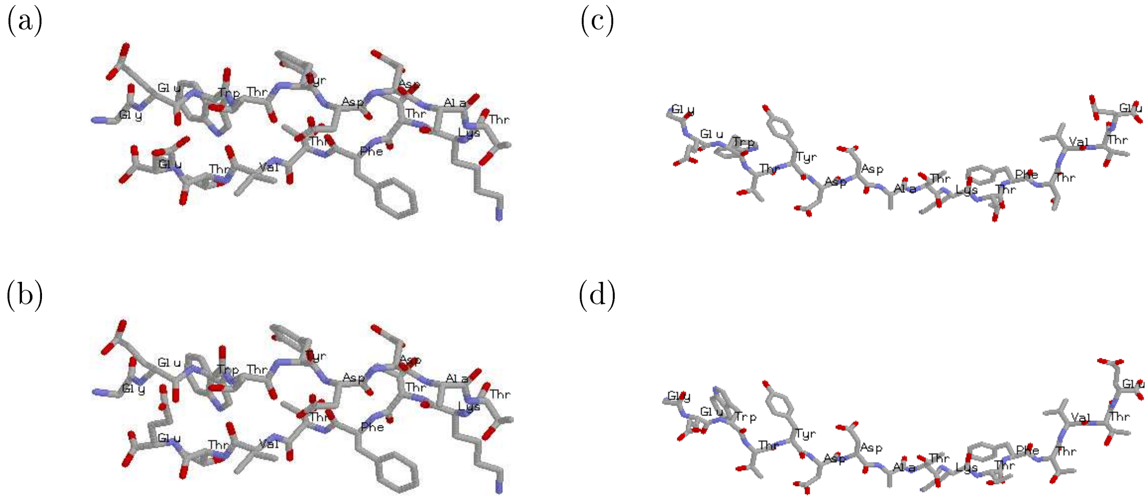

Figure 52: (a) Overall minimum energy conformation ($E = 653.020$, $F = 602.768$). (b) Overall minimum free energy conformation ($E = 654.139$, $F = 602.647$). (c) Minimum energy extended conformation ($E = 673.439$, $F = 605.342$). (d) Minimum free energy extended conformation ($E = 675.854$, $F = 604.347$). Energy and free energy values are expressed in kcal/mole. Free energy values are at $T = 300\,\text{K}$.

In order to study the folding pathways of Protein G (41–56), we need to generate an adequate sample of stationary points of the potential energy surface. Not only do we need to generate conformations which resemble the hairpin native state, as well as extended conformations, but we also need to find conformations which lie along the low-lying pathways connecting these two regions of conformation space. Thus, we need to find low-lying conformations, as well as transition states, over a large region of conformation space.

The approach we have chosen to follow is to first generate an initial sampling of minima, forming a "scaffolding", and then building upon that scaffolding by performing uphill and downhill searches using an eigenmode-following algorithm. Before carrying out this search, however, we first want to identify the global minimum energy conformation on the potential energy surface, which will serve as the native structure. The results of the global minimum search are given in Figure 52. The study of the pathways for the transitions from extended to beta-sheet conformations is currently in progress.

V Protein-Protein Interactions

Understanding protein-protein interactions, also known as peptide docking, is critically important for rational protein engineering and pharmaceutical design. Peptide docking is the binding of one protein to another protein, and such binding is essential to processes ranging from chemotherapy to the communications between cells. Advances in understanding and predicting how solvation, electrostatics, and other forces affect the strength, specificity, and kinetics of peptide docking interactions is vital for discovering new drugs, for developing tools for characterizing and treating disease, and for designing sensors and other molecular recognition devices. No comprehensive peptide docking prediction method yet exists. In this section, we review the state of the art in peptide docking prediction methods.

A Background

Predicting peptide docking and protein-protein interactions computationally involves predicting the shapes, characteristics, and interactions of "target" molecules and the "docking" ligand molecules which bind to them. One part of the prediction challenge involves determining the conformation or structure of the binding sites in the target molecule. The other part of the prediction challenge involves determining the binding affinity of different docking molecules for the target molecule. This includes identifying a set of equilibrium structures for complexes between different docking molecules and the target molecule, and then quantifying and comparing, or "scoring", the binding affinity of docking molecule structures. The following two sections discuss methods used for binding site structure prediction and binding affinity prediction.

1 Prediction of Binding Site Structure

The identification of binding site conformations in target molecules usually requires experimental structure determination of the binding site. One class of protiens that has received particular attention is the class of proteins derived from the major histocompatibility complex (MHC), a set of genes critical in the immune response [230]. Crystallographic studies have been performed for the two major classes of MHC molecules, class I [74, 215] and II [221]. Such crystallographic information is invaluable. For instance, it can define rigid binding sites for docking molecules and thus greatly reduce the conformational space being searched in computational searches for structures of target/docking molecule complexes.

The determination of high quality models of protein structure for which no experimentally determined coordinates exist has received considerable attention in the literature. A commonly used approach is based on homology modeling, in which a model for a target protein is generated using the known structure of a homologous protein. Typcially a backbone model first is constructed for the structurally conserved regions, and then loops and side chains are added [27, 226]. For the prediction of side-chain conformations, many approaches based on homology modeling are available. These approaches differ from each other in (a) the rotamer libraries used, (b) the energy function chosen, and (c) the search strategy employed. In sampling conformational space through rotamer libraries, many different approaches have been used, including backbone-independent rotamer libraries [41] or rotamer sets that incorporate backbone-sidechain interactions [63]. Also employed are extended rotamer libraries derived from cluster analyses of experimentally determined databases [204], as well as augmented libraries that use discrete values around observed χ angle values $\pm 10°$ [233]. Regarding the energy function used, simplistic local interactions typically are limited to van der Waals or hard-sphere energies [43, 133]. Finally, the employed search strategies are mainly heuristic methods involving Monte Carlo techniques [99], genetic algorithms [229], neural networks [105], mean-field optimization [133], and combinatorial searches [63].

Recently, a novel decomposition-based approach has been proposed for predicting binding site structures in the MHC II HLA-DR1 protein [106]. In this approach, existing MHC II crystal structures are used to predict the binding site conformations of other MHC II molecules. The approach uses the detailed potential energy force field ECEPP/3 and an area-based solvation method. A global optimization search, based on the αBB algorithm, is used to identify the global minimum energy conformation of the binding sites. As discussed further in later sections, the predicted binding sites agree with available crystallographic data with only small rms deviations.

2 Prediction of Binding Affinity

The development of accurate "scoring" functions to identify and compare equilibrium structures of target/docking molecule complexes is a challenging and unsolved problem. A general scoring function is represented in Equation 101:

$$\Delta G = \Delta G_{complex} - \Delta G_{ligand} - \Delta G_{pocket} \tag{101}$$

Here $\Delta G_{complex}$, ΔG_{binder} and ΔG_{pocket} are the free energies of the target/docking molecule complex, the free docking ligand molecule, and the free target pocket or binding site, respectively. ΔG is then the free energy of binding or binding affinity.

Due to the computational complexity of rigorous energy calculations, many methods have relied on qualitative modeling of peptide docking interactions. As a first approximation, models have been developed which assume the docking and target molecules are rigid. In this rigid binding approximation case, the use of shape complementarity has had some limited success [136]. Such algorithms model the ligand and target macromolecule according to their surface topology, and attempt to identify which complexes exhibit the best "fit". Here scoring functions are based on the complementarity of the molecules, which, in most cases, is related to their solvent accessible surface areas [46, 47, 142]. The strength of these methods is that they can be made computationally efficient and used to screen large databases of potential ligands. However, studies comparing the computational results of these methods to experimentally determined native complexes indicate that rigid models identify many non-native low energy structures. The rigid-docking scoring function can be refined by adding additional components, such as conformational energy and solvation energy.

On the other end of the flexibility spectrum are fully flexibile, exact models. It has been demonstrated that exact modeling of binding free energies provides results in nearly exact quantitative agreement with experimental results [89, 113, 211]. In contrast to the rigid description of docking, these methods allow for flexibility of both the ligand and receptor molecules. However, for general peptide docking problems, these thermodynamic integration and free energy perturbation methods are computationally infeasible with current computing power. These problems are only tractable when approximate structures are known and relatively small. More detail on these methods can be found elsewhere [170, 197]. A comprehensive theoretical treatment of the thermodynamics of binding processes in macromolecules is also available [38].

More computationally feasible methods are based on calculating binding free energies using empirically derived free energy functions. Some methods of approximating free energy functions involve structure-based potentials [239]. Other approximations utilize parameterization of experimental data to construct scoring functions based on conformational energy, hydrophobic and hydrophilic surface areas, and hydrogen bonding geometries [97, 235]. However, these methods are generally nontransferable.

A more universal approach, applicable to flexible ligands, is to base free energy calculations on general force field models, which involve potential energy functions similar to those described in the preceding sections. This free energy function must also account for solvation energy, which can be calculated from structure based solvation terms or continuum-based models of solvation. Rigorously, entropic effects of side chain rotations should also be considered. Reviews of methods used to evaluate binding free energies can be found elsewhere [108, 231].

Once a method for "scoring" the binding affinity has been selected, the exact form of the approach for determining and optimizing the target/docking molecule complex must be developed. Several general approaches have been employed. The most obvious and most difficult approach would be to optimize the entire system of the two interacting peptides. To accomplish this, the relative position of the two peptides, which is defined by six degrees of freedom (three translation and three rotation), along with the total number of internal degrees of freedom for the two molecules, must be considered. This problem becomes intractable for all but the smallest systems. Alternative approaches have decomposed the problem by considering the binding affinities of shorter subsequences at different binding sites of the target macromolecule. The full binding ligand can then be constructed based on the optimally docked subsequences. This approach relies on the ability to build a suitable ligand. Another alternative method is based on independently generating conformations of the isolated ligand. Binding affinities for a number of these rigid conformations then can be calculated and compared, with the drawback that conformations with higher binding affinities may be overlooked.

The following discussion classifies peptide docking approaches according to their treatment of the internal flexibility of the docking ligand molecule. Some approaches combine aspects of both rigid and flexible methods, and the choice of scoring function is often closely related to these classifications. For example, it is implicitly difficult for shape-based approaches to capture internal flexibility due to their simplified description of the molecular surface. Detailed energy-based approaches better represent the free energy of the system and can represent internal conformational changes, but their increased dimensionality make these methods more computationally expensive. The complexity of these approaches indicates that rigorous global optimization methods are needed to address the peptide docking prediction challenge.

Rigid Models The first, and most common, methods used to address the peptide docking problems were based on the concept of shape complementarity. These methods employ, at least initially, rigid approximations for both the docking ligand and target receptor molecules. In the most general case six degrees of freedom, three translational and three rotational, must be optimized to determine the best "fit" for the receptor-ligand complex. Approximations often are used in practice to reduce the number of degrees of freedom. In addition, the alignment of each ligand must be optimized within the binding site. Typically, several screening stages are used to reduce these optimizations to a manageable number.

One shape-based method utilizes a simplified protein model, which is generated by representing each amino acid by a single sphere. The scoring function is based on interfacial areas and a simplified nonbonded potential energy term. Potential ligand structures are screened by systematically rotating the ligand and then translating the structure, along only one dimension, into the pocket [109, 250, 249]. These approximations and simplifications are necessary in order to make the problem tractable, especially in the context of a systematic search. A recent modification attempts to overcome these computational limitations by using a simulated annealing, rather than a systematic search, to screen the ligand structures [42].

Distinctive characteristics of molecular surfaces also have been used to reduce the number of degrees of freedom for shape-based docking problems. One study considers local shape functions, which are generated by placing spheres at surface points along the docking ligand and target receptor surfaces. The volume within the surface and the unit vector that extends from the center of the sphere to the surface characterize these functions. A combinatorial algorithm can then be used to compare these local shape functions at "knobs and holes" [146] on the ligand and receptor surfaces so that the best alignments of the two molecules can be identified [48].

More detailed descriptions of molecular surfaces also have been used in determining shape complementarity. One procedure creates a webbed surface for the ligand and receptor by using a local coordinate system to define the surface points for each molecule. After setting the ligand position, a least squares method is used to align the surface points of the two molecules. The method also screens ligands according to a Coulombic scoring function [15].

An alternative approach transforms the problem from identifying complementary shapes for the receptor and ligand proteins into one of matching similar shapes for these two molecules. This is accomplished by describing the binding site as a collection of spheres that lie on the outside of the receptor surface, and characterizing the ligand as a collection of spheres that lie on the inside of the ligand surface [56, 57, 136]. Potential matches are identified by grouping and comparing distances between the center of spheres for each molecule. Local refinement of translation and rotation vectors is used for the highest ranking matches. The complexity of the problem is to some degree obscured, because it also depends on the choice of location, size and number of spheres used to model the receptor molecule. Other modifications of this procedure include the addition of hydrogen bonding criteria and the use of local minimization of the potential energy in order to relax the rigidity of the ligand molecule [141, 213].

The "soft docking" model represents the target and docking molecules as a collection of cubes rather than spheres. This method combines aspects of surface complementarity, grid search and soft potential modeling. The "cubic" representation along with a grid search makes the translational and rotational searches much more efficient. In addition, the cubes implicitly allow for some volume overlap, which can be used in combination with surface complementarity to screen docked complexes [111].

In general, when considering a rigid receptor, the concept of a grid search can be used to reduce the computational requirements of evaluating scoring functions. This is accomplished by precomputing values for the receptor based on points of a three dimensional grid [85]. The concept is similar to cubic lattice model approaches in molecular conformation problems, for which a recently proposed algorithm using a tabu search has been highly effective [190]. This approach has been the basis of a number of recent studies [167, 240], including one that employs a Monte Carlo search in the context of "knobs and holes" docking [240].

Flexible Models In the most general case, flexible docking approaches attempt to optimize the free energy of the entire target/docking mlecule complex, which is described by translational, rotational and internal variables of the system. In contrast to most rigid modeling approaches, these methods typically do not require prior knowledge of ligand conformations. As a result, their success in predicting ligand binding is highly dependent on the use of detailed scoring functions to evaluate free energy changes. In addition, although some studies have considered full macromolecular-ligand systems, most approaches also depend on effective decomposition strategies of the overall docking problem.

Several simple approaches have been implemented in an attempt to model flexible docking. For example, a number of methods have incorporated ligand flexibility by considering databases of multiple ligand conformations [124, 168]. However, these methods require reliable databases and methods for developing appropriate ligand conformations, and these are typically not available. On the other hand, thermodynamic integration and free energy perturbation methods allow for full flexibility and detailed modeling of binding free energies. However, these simulations, usually accomplished by molecular dynamics, effectively explore only single low energy minima. This has led to the need for global optimization methods that efficiently search the conformational energy hypersurface associated with peptide docking problems.

One of the most common approaches is based on Monte Carlo simulating annealing algorithms. This method was first applied to flexible ligand docking using molecular affinity potentials [86]. Molecular affinity potentials increase the computational efficiency of the search by employing precomputed energy grids [85]. In this case, flexibility is introduced by allowing internal rotations of torsion angles, along with translational and rotational movement. However, for each docking example, a set of simulated annealing runs is necessary in order to increase the confidence of the reported structures.

A second method, also based on simulated annealing, involves a two-step procedure to dock flexible oligopeptide ligands [36]. In the first step, a modified potential energy force field is used to reduce unfavorable intermolecular contacts. This energy model is employed in local energy minimizations of arbitrarily docked ligands, which are needed in order to generate an initial set of ligand conformations. The scoring function for the second step describes energy interactions between the flexible ligand and rigid receptor molecules. The set of minimized conformations is then used to generate starting points for a Monte Carlo minimization procedure. Although experimental results were not initially available, later comparison has shown that this method does not correctly predict MHC binding. These discrepancies are most likely attributable to incorrect energy modeling (e.g., no inclusion of solvation), along with the inherent inefficiencies associated with simulated annealing searches.

Another MC-based method employs a multiple-start technique in an attempt to reproduce the results of a systematic search. The first step involves a Monte Carlo search with a grid-based scoring function in order to limit steric overlaps of the ligand and receptor molecules. A second, energy-directed, simulated annealing search uses a pairwise potential energy function. Rather than rely on a single search, this method employs a large number of short simulated annealing runs. Although initial results were based on both rigid receptor and ligand conformations [95], more recent work has addressed the issue of flexible ligand docking [96].

Another type of MC method is the scaled collective variable Monte Carlo method used in the software package PRODOCK [227]. This method performs energy minimizations after each MC step, which helps to distinguish native conformations from low-energy non-native conformations. Bezier splines and other techniques have been incorporated into the method to improve its efficiency. In addition, PRODOCK allows different amino acids in the docking complex to be defined as rigid or flexible.

In a similar way, genetic algorithms recently have been used to dock flexible ligands. In some cases, scoring functions have been based on potential energy force fields [188], although some modified potentials also have been used [112]. The results of one method [115], which includes solvation effects, emphasize the need for developing reliable scoring functions. In general, as with simulated annealing, the ability to model flexibility is limited as ligand size increases. The coupling of these effects with the implicit unreliability of both the genetic algorithm and simulated annealing search techniques must be closely considered when approaching large scale docking problems such as de novo drug design.

Combinatorial methods also have been used to address the difficulties of modeling full ligand flexibility. In theory, these methods are similar to build-up methods used for the protein folding problem, although in peptide docking intermolecular interactions also are included. An initial application to the peptide docking problem was based on rigid ligand models generated from a database of chemical structures [57]. A more detailed implementation uses libraries of low energy conformations for single amino acid residues. These conformations are subsequently joined and grouped according to scoring functions based on the intra- and inter-molecular energies of the target/docking ligand complex [176]. More recent methods have employed databases developed for smaller ligand fragments such as functional groups [202] or even atoms [201]. In general, these ligand build-ups are initialized by selecting a starting point within the target binding site pocket. As with the protein folding approaches, such combinatorial techniques must employ effective reduction schemes in order to limit the number of generated conformations.

Similar approaches combine the ideas of fragment assembly and site mapping. In contrast to the single anchor requirement of simple build-up methods, these techniques attempt to identify a number of anchor fragments or residues which can be joined through a process of fragment assembly. The first step, site mapping, is equivalent to docking probe fragments at specific sites of the target macromolecule. Some methods have screened the binding affinities of these probes using shape-based modeling [139], while others have relied on other energy-based descriptions, such as hydrogen bonding interactions [28, 29]. In general these site maps are constructed by local minimization, grid or library searches of the probe conformations. Other techniques employ a multiple copy simultaneous search [169, 200]. Once anchor positions have been determined using one of these methods, the resulting segments must be joined by fragment assembly. Bridges can be formed by searching through molecular libraries, or in some cases using an exhaustive search over all connections [35]. A recently proposed technique applies a dynamic programming approach, as discussed above, to the fragment assembly phase of a nonameric ligand in an MHC HLA-A2 complex [90]. A molecular dynamics simulation also has been utilized for studying the binding afinity of the HLA-B*2705 protein [199].

Recently, a novel decomposition-based approach has been proposed for predicting binding site structure and peptide docking to the MHC II HLA-DR1 protein [12]. The approach performs site mappings of the five polymorphic pockets of MHC II molecules that accommodate peptide docking [221]. In one part of the approach, existing MHC II crystal structures are used to predict the binding site conformations of other MHC II molecules. In another part of the approach, each naturally occurring amino acid is treated as a probe molecule for each of the five pockets. The approach uses a deterministic global optimization search technique to identify the best conformation for each pocket or residue. The scoring function accounts for both intra- and inter-molecular interactions using the detailed potential energy force field ECEPP/3 along with several solvation model approaches. The global optimization search, based on the αBB algorithm, is used to identify the global minimum energy conformation for the pockets and for both the bound and free residues. The corresponding energy differences are then used to provide rank ordered lists of the best binders for each pocket. As discussed in later sections, results for pocket 1 of the HLA-DRB1 macromolecule have exhibited good agreement with experimental binding assays [12].

A recent review of approaches for peptide docking can be found in Floudas et al. [72]. The main disadvantages of most of these approaches are:

(a) only a very limited conformational space is considered, since usually fewer than 10 rotamers are used for each residue,

(b) the simplicity of the energy functions is not able to give realistic description of the molecular system, and

(c) no systematic search methodology exists to guarantee the determination of the global optimal solution, even in methods using simplified energy functions.

Thus many current models of binding site structure prediction and binding affinity prediction in peptide docking are not able to guarantee that they have found the optimum docking solution because they consider only a few of the many conformations two docking partners may adopt, because they are not quantitative, or because they do not fully consider entropic, electrostatic, or other energetic effects.

B Prediction of Binding Site Structure

We have developed a theoretical approach that, based on crystallographic data from MHC II molecules, determines the three-dimensional structure of MHC II molecule binding sites for which crystallographic data is not available. Class II histocompatibility molecules are cell surface molecules that form complexes with self and non-self peptides and then present them to T-cells as part of the immune response. A number of class II histocompatibility molecules have been analyzed by crystallography, including HLA-DR1 [221], HLA-DR3 [78], and I-E^k [73]. Crystal structures are not available, however, for the vast majority of class II MHC molecules. MHC II molecules for which crystal structures are not available are important in autoimmune diseases such as diabetes and rheumatoid arthritis, and being able to predict such structures would advance the understanding and treatment of these diseases.

Our approach to binding site structure prediction uses the ECEPP/3 potential energy model for describing the energetics of atomic interactions (as described in Section III.A.1 above) and employs the rigorous deterministic global optimization algorithm αBB (as described in Section II.A.6 above) to obtain the global minimum energy conformation of the binding site. With this approach, we predicted the binding sites of HLA-DR3 and I-E^k based on the crystallographic structure of HLA-DR1 [221]. The root mean square differences (based on all atoms) between the structures we predicted and the actual crystal structures of the two molecules [73, 78] are between 1.09-2.03$\mathring{A}$. We also calculated the binding affinity of our predicted structures using the decomposition approach discussed in Section V.C.2. These binding affinities are in good agreement with the results obtained by applying the decomposition approach to the actual crystal structures.

1 Definition of Problem

The recent crystallographic studies of class II HLA molecules [221, 78, 73] suggest an overall similarity in their structures. The conformation of HLA-DR3 in the HLA-DR3-CLIP complex is only slightly different from that of HLA-DR1 in HLA-DR1-HA [78], and a comparison of two I-E^k structures with HLA-DR1 identifies that only a few differences in β chain amino acids exist between I-E^k and both the HLA-DR1 and HLA-DR3 sequences. However, these few variable residues are sufficient to explain antigenic differences without recourse to allosteric transitions or novel conformations.

Consequently, specific information about the structure of the histocompatibility molecules is needed in order to be able to analyze their specificity. Because crystal structures of class II molecules are not available except for the human crystals of HLA-DR1-HA and HLA-DR3-CLIP and the murine crystals I-E^k-HB, I-E^k-Hsp, we propose a novel approach based on decomposition and deterministic global optimization that enables the prediction of the three dimensional structure of the binding sites of class II molecules and can be used efficiently for the qualitative assessment of their binding affinities.

The question that is addressed is stated as follows:

"Given the (x,y,z) coordinates of the atoms in pockets 1,4,6,7 and 9 of the HLA-DR1

[221], can we predict the three dimensional structures of the corresponding pockets of HLA-DR3 and I-E^k?"

2 Approach

A systematic approach is presented below for the structure prediction of an antigen binding site based on the crystallographic data of the HLA-DR1 molecule [221]. The approach examines each of the binding sites separately and involves the following steps:

(1) The binding sites of HLA-DR1 molecule are evaluated. All amino acids within a radius of $\mathcal{R}$=5.0$\mathring{A}$ of the atoms of the binding amino acid in the crystallographic studies [221] are identified as shown in Table 36. The Program for Pocket Definition, as described in [12] and Section V.C.3, constructs these pockets through the selection of all residues that are within a radius $\mathcal{R}$ of the atoms of the crystallographic binder.

(2) The amino acid substitutions between HLA-DR1 and the HLA-II molecule (e.g., HLA-DR3, I-E^k) are identified and are shown in Table 37. Note that pocket 1 of HLA-DR1 requires only one substitution Gly$\rightarrow$ Val in position β86 give pocket 1 of HLA-DR3. Pockets 4, 6 and 7 involve three substitutions while pocket 9 features only one substitution in the representation of the corresponding pockets of HLA- DR3. Note also that all pockets of HLA-DR1 require three or four substitutions in order to give the corresponding pockets of I-E^k.

(3) For each one of the substituted residues, the intra and inter-molecular energy interactions are modeled. Specifically, the electrostatic, nonbonded, torsional, and hydrogen bonding contributions [180], are considered for the substituted residue, as well as the interactions of the substituted residues with the rest of the amino acids that constitute the examined binding site. The solvation energy also is considered through solvent-accessible areas [191, 236] as explained in Section III.A.2. The dihedral angles that define the three-dimensional structure of the substituted residues are considered explicitly as variables. The relative position of each amino acid also must be determined, and this is done through the determination of each amino acid's translation vector and Euler angles. For the substituted amino acids lower and upper bounds are considered for their N and C' coordinates, based on the available crystallographic data [221, 78, 73].

(4) Having the mathematical model that includes the intra- and inter-molecular energetic interactions and the solvation energy, and which has as variables the dihedral angles of the substituted amino acids as well as their translation vectors and Euler angles, we minimize the total potential energy by employing the α**BB** deterministic global optimization approach [5, 9, 3, 4, 2] as described in later sections below.

(5) The resulting global minimum energy conformer provides information on the predicted (x,y,z) coordinates of the atoms of the substituted residues. Structure verification is made by superposition of all atoms of the predicted structure and the ones derived from crystallographic

Table 36: HLA-DR1 Pocket Compositions for $\mathcal{R} = 5.0$ Å

Pocket				
1	**4**	**6**	**7**	**9**
pheα24	glnα09	gluα11	valα65	asnα69
ileα31	gluα11	asnα62	asnα69	leuα70
pheα32	asnα62	valα65	gluβ28	ileα72
trpα43	pheβ13	aspα66	tyrβ47	metα73
alaα52	leuβ26	leuβ11	trpβ61	argα76
serα53	glnβ70	pheβ13	leuβ67	trpβ09
pheα54	argβ71	argβ71	argβ71	aspβ57
gluα55	alaβ74			tyrβ60
asnβ82	tyrβ78			trpβ61
valβ85				
glyβ86				
pheβ89				
thrβ90				

data. The superposition is based on the global minimization of the root mean square differences of the distances between all the atoms involved in the pocket as described in the Computational Studies section below.

3 Modeling

When bond angles and bond lengths are assumed to be rigid, the geometric shape of a protein is uniquely determined by its dihedral angles. If more than one polypeptide is involved, the relative orientations and locations of these different chains also must be defined. This can most easily be accomplished by defining a translation vector and a rotation matrix. The translation vector is based on the Cartesian coordinates of the initial nitrogen atom of each independent chain. Euler angles specify the rotations necessary to orient a particular polypeptide and are defined as the angles between the coordinate axes defined by the initial hydrogen, nitrogen, and alpha carbon of each residue.

The system under study involves all the residues of the binding site. The substituted amino acids constitute the problem variables, whereas the residues that remain the same are treated as fixed based on the crystallographic data. Since there may be multiple amino acid substitutions, the problem variables are the amino coordinates, the Euler angles, and the dihedral angles of all substituted residues. In contrast to other existing approaches, the Euler angles and dihedral angles are considered to span the entire feasible range [-180°, +180°] and are not restricted to specified discrete values.

Consequently, the total energy function is defined as:

$$E_{Total} = E_{Unsolvated}^{MIN} + E_{Solvated} \qquad (102)$$

Table 37: Substitutions for HLA-DR3 and I-E^k Binding Sites

Pocket	Substitutions for HLA-DR3	Substitutions for I-E^k
1	β86: Gly $\rightarrow$ Val	β85: Val $\rightarrow$ Ile β86: Gly $\rightarrow$ Phe β90: Thr $\rightarrow$ Leu
4	β13: Phe $\rightarrow$ Ser β26: Leu $\rightarrow$ Tyr β74: Ala $\rightarrow$ Arg	β13: Phe $\rightarrow$ Ser β74: Ala $\rightarrow$ Glu β78: Tyr $\rightarrow$ Val β:71 Arg $\rightarrow$ Lys
6	β11: Leu $\rightarrow$ Ser β13: Phe $\rightarrow$ Ser β71: Arg $\rightarrow$ Lys	β11: Leu $\rightarrow$ Ser β13: Phe $\rightarrow$ Cys β71: Arg $\rightarrow$ Lys
7	β28: Glu $\rightarrow$ Asp β47: Tyr $\rightarrow$ Phe β71: Arg $\rightarrow$ Lys	β28: Glu $\rightarrow$ Val β47: Tyr $\rightarrow$ Phe β67: Leu $\rightarrow$ Phe β71: Arg $\rightarrow$ Lys
9	β9: Trp $\rightarrow$ Glu	α72: Ile $\rightarrow$ Val β9: Trp $\rightarrow$ Glu β60: Tyr $\rightarrow$ Asn

where $E^{MIN}_{Unsolvated}$ is the potential energy of the system without considering solvation, $E_{Solvated}$ is the solvation energy of the system, and E_{Total} is the potential and solvation energy of the system. Based on the above description the mathematical formulation can be posed in the following way:

$$\min \quad E_{Total}\left(\phi^m, \psi^m, \omega^m, \chi_k^m, N_x^m, N_y^m, N_z^m, \varepsilon_1^m, \varepsilon_2^m, \varepsilon_3^m\right) \tag{103}$$

$$\begin{aligned}
\text{s.t.} \quad -\pi &\leq \phi^m, \psi^m, \omega^m, \chi_k^m, \varepsilon_1^m, \varepsilon_2^m, \varepsilon_3^m \leq \pi && (104)\\
(N_x^m)^L &\leq N_x^m \leq (N_x^m)^U && (105)\\
(N_y^m)^L &\leq N_y^m \leq (N_y^m)^U && (106)\\
(N_z^m)^L &\leq N_z^m \leq (N_z^m)^U && (107)
\end{aligned}$$

$$\begin{aligned}
(C_x'^m)^L &\leq C_x'^m\left(\phi^m, \psi^m, \omega^m, \chi_k^m, N_x^m, N_y^m, N_z^m, \varepsilon_1^m, \varepsilon_2^m, \varepsilon_3^m\right) \leq (C_x'^m)^U && (108)\\
(C_y'^m)^L &\leq C_y'^m\left(\phi^m, \psi^m, \omega^m, \chi_k^m, N_x^m, N_y^m, N_z^m, \varepsilon_1^m, \varepsilon_2^m, \varepsilon_3^m\right) \leq (C_y'^m)^U && (109)\\
(C_z'^m)^L &\leq C_z'^m\left(\phi^m, \psi^m, \omega^m, \chi_k^m, N_x^m, N_y^m, N_z^m, \varepsilon_1^m, \varepsilon_2^m, \varepsilon_3^m\right) \leq (C_z'^m)^U && (110)
\end{aligned}$$

where m=1,...,M corresponds to total number of substitutions.

The additional constraints (105-110) represent the bounds on the N and C' coordinates and express the binding of the specific residue with the rest of pocket [12], since the substituted residue is part of a longer polypeptide and consequently is not allowed to rotate freely. Since the C' coordinates can be evaluated as functions of the independent variables, the restrictions on the position of C' are implemented by the incorporation of a penalty function, P, in the objective function.

$$
\begin{aligned}
P \ = \beta\{ \quad & \langle C_x^{\prime l} - C_x^{\prime} \rangle + \langle C_x^{\prime} - C_x^{\prime u} \rangle + \\
& \langle C_y^{\prime l} - C_y^{\prime} \rangle + \langle C_y^{\prime} - C_y^{\prime u} \rangle + \\
& \langle C_z^{\prime l} - C_z^{\prime} \rangle + \langle C_z^{\prime} - C_z^{\prime u} \rangle \ \}
\end{aligned}
$$

The $\langle\rangle$ function is defined as follows: $\langle\mathcal{A}\rangle$ equals $\mathcal{A}$ if $\mathcal{A}$ is greater than zero, otherwise $\langle\mathcal{A}\rangle$ equals zero. Thus, any coordinate value beyond the specified bounds is multiplied by the penalty parameter β and added to the potential energy. Consequently, the minimization of the objective function eliminates solutions in which the C' position falls outside the specified bounds.

The global optimization formulation is then as follows:

$$
\begin{aligned}
L \ = \ E_{Total} + \alpha \{ \sum_{m=1}^{M} \left(\phi^{mL} - \phi^{m}\right)\left(\phi^{mU} - \phi^{m}\right) + \left(\psi^{mL} - \psi^{m}\right)\left(\psi^{mU} - \psi^{m}\right) + \\
\left(\omega^{mL} - \omega^{m}\right)\left(\omega^{mU} - \omega^{m}\right) + \sum_{k=1}^{K} \left(\chi_k^{mL} - \chi_k^{m}\right)\left(\chi_k^{mU} - \chi_k^{m}\right) + \left(N_x^{mL} - N_x^{m}\right)\left(N_x^{mU} - N_x^{m}\right) + \\
\left(N_y^{mL} - N_y^{m}\right)\left(N_y^{mU} - N_y^{m}\right) + \left(N_z^{mL} - N_z^{m}\right)\left(N_z^{mU} - N_z^{m}\right) + \left(\varepsilon_1^{mL} - \varepsilon_1^{m}\right)\left(\varepsilon_1^{mU} - \varepsilon_1^{m}\right) + \\
\left(\varepsilon_2^{mL} - \varepsilon_2^{m}\right)\left(\varepsilon_2^{mU} - \varepsilon_2^{m}\right) + \left(\varepsilon_3^{mL} - \varepsilon_3^{m}\right)\left(\varepsilon_3^{mU} - \varepsilon_3^{m}\right) \ \}
\end{aligned}
$$

where α is a nonnegative parameter which must be greater or equal to the negative one half of the minimum eigenvalue of the Hessian of E_{Total} in the considered domain defined by the lower and upper bounds (i.e., $x^L = -\pi, x^U = \pi$) of the dihedral angles, translation variables and Euler angles. This parameter can be rigorously calculated based on the techniques introduces by Adjiman and Floudas [5] and Adjiman et al. [4, 2].

For the problem of determining the binding sites of the unknown HLA molecules, the global variable set includes the ϕ, ψ and χ_1 variables. All the dihedral angles of the substituted residues, as well as their translation vectors and Euler angles, are continuous variables in the problem and are treated as local variables.

4 Deterministic Global Optimization

The implementation of the approach involves the connection of the conformational energy program PACK [205], which allows the evaluation of all energy interactions when more than one protein chain is involved in the system, to the deterministic global optimization framework αBB. PACK evaluates all energy components through repeated calls to the ECEPP/3 potential energy function program. The local optimization solver NPSOL is used for the minimization of the overall potential energy provided by PACK and for the minimization of the convexified potential function (L) provided by αBB. MSEED [191], the program for the determination of solvation energy, is interfaced to αBB to allow the consideration of the solvation energy at the local minima. The algorithmic procedure is represented graphically in Figure 53.

The implementation of the proposed approach is illustrated in Figure 54 and involves the following steps:

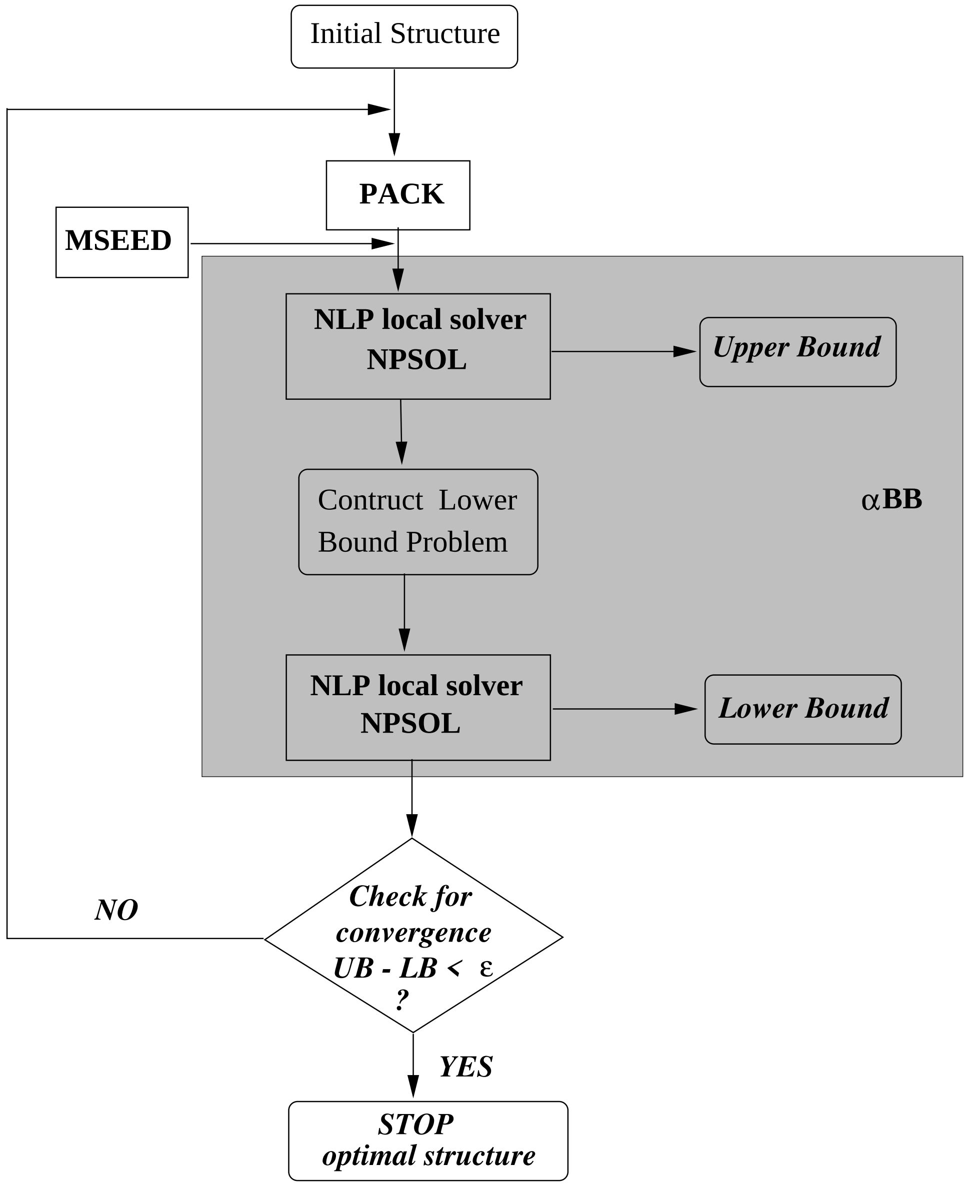

Figure 53: Deterministic Global Optimization Algorithm for Binding Site Structure Prediction

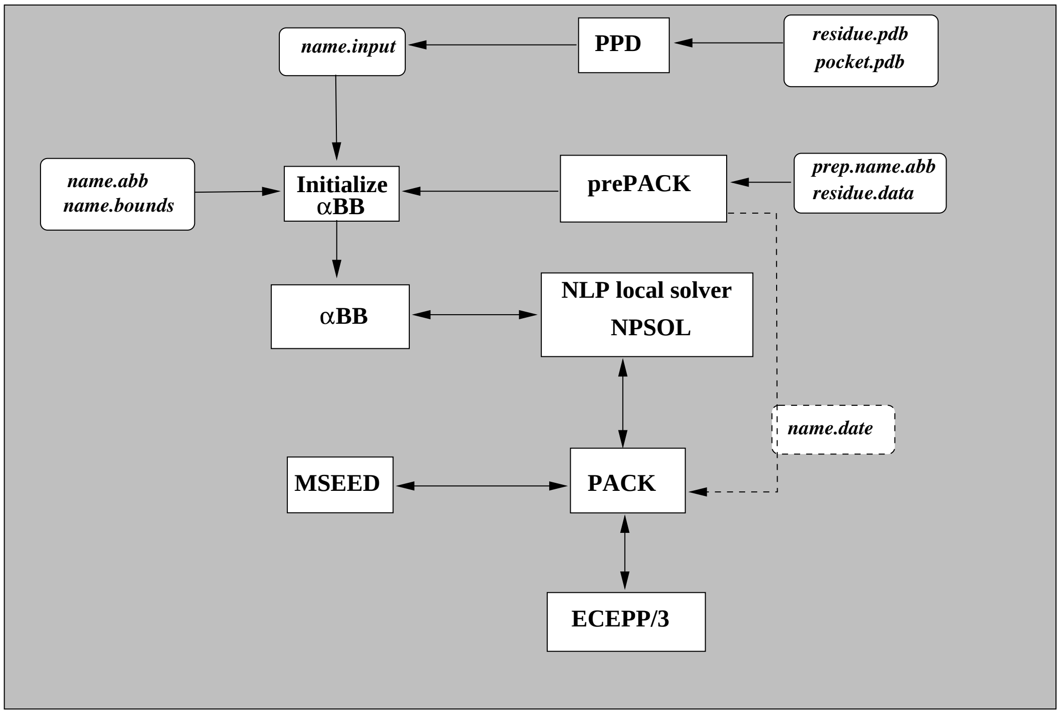

Figure 54: Implementation of the Binding Site Structure Prediction Approach

(1) The program for Pocket Definition, (*PPD*), uses the input files *residue.pdb* and *pocket.pdb* to generate the coordinates of the residues involved in the considered pocket.

(2) The program *ARAS* is used to determine the translation vectors, Euler angles and dihedral angles of the residues in the pocket given their (x, y, z) coordinates. This information and the initial values for the translation vector, Euler angles, and dihedral angles of the substituted residues are incorporated within the input file *name.input*.

(3) The program *prePACK* utilizes the *residue.data* file (a set of initial atomic coordinates that are based on fixed bond lengths, fixed bond angles and each variable dihedral angle initially set to 180°), the *mol.in* file for each one of the amino acids involved in the pocket, and the *prep.name.abb* file (which specifies the fixed and substituted residues) to create a *name.date* file. The *name.date* file is the standard input for the potential function program, *PACK*.

(4) The global optimization program αBB requires a *name.abb* file that defines the optimization problem, including the variable bounds. αBB also uses the *name.input* file and the *name.bounds* file, which contains the C' bounds used to evaluate the coordinates of C' as a function of the independent variables.

(5) The program *PACK* uses the *name.date* file and is connected with *ECEPP/3* in or-

der to evaluate the potential function, which is minimized by the local optimization solver *NPSOL*.

(6) The *MSEED* solvation energy program uses the *JRF.dat* file which defines the solvation parameters σ_i and evaluates the solvation energy at the current local minimum structure.

5 Computational Studies

Comparison with Crystallographic Data To compare the predicted structure of the pockets accurately with the crystallographic data, the best rotation and translation to relate the two different sets of atomic position must be obtained. Given two proteins A and B with N_{atom} atoms, the best superposition is the one that minimizes the sum of squared distances between each B atom and the corresponding A atom. Existing approaches to this problem are based on:

(i) iterative minimization using rotation angles [196, 194];

(ii) the use of decomposition approaches, where the transformation matrix L is determined by calculating the best unrestricted linear transformation which converts A into B using the least squares matrix method [58]; or the formation of a generalized inverse of the molecular structure [155], and then the decomposition L=RS where R is a rotation matrix and S is a symmetric distortion matrix;

(iii) the construction of a matrix U which yields an orthogonal rotation directly [118, 119, 163, 164].

As pointed out by McLachlan [164], the rotation angles method is very slow, while the rotation matrix methods depend on whether A is fitted to B or vice versa and do not minimize the r.m.s distance. McLachlan [164] proposed an approach to improve the speed and accuracy of determining the matrix U and moreover to cover all special cases which arise when U is degenerate or singular.

We formulated and solved the problem of obtaining the best fit of two protein structures as a global optimization problem. The best rotation and translation matrices that minimize the "fitting distance" for the two protein structures are guaranteed to be found in all special cases without having to perform any additional tests and calculations.

Consider two protein structures A and B, with A obtained from the crystallographic data and B determined from our methodology. Both structures involve N_{atom} atoms with Cartesian coordinates $(x_c(i), y_c(i), z_c(i))$ for the crystal and $(x_p(i), y_p(i), z_p(i))$ for the predicted structure. The mathematical formulation of the best fitting problem can then be posed as follows:

$$\min \quad RMS = (1/N_{atom})\sqrt{\sum_{i=1}^{N_{atom}} (x_c(i) - x'(i))^2 + (y_c(i) - y'(i))^2 + (z_c(i) - z'(i))^2}$$

s.t.

$$\begin{bmatrix} x'(i) \\ y'(i) \\ z'(i) \end{bmatrix} = R \begin{bmatrix} x_p(i) \\ y_p(i) \\ z_p(i) \end{bmatrix} + T$$

$$
R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{22} & r_{22} & r_{22} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}, \quad T = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix}
$$

$$
RR^\top = I \qquad (111)
$$

where R and T are the required rotation and translation vectors that translate the predicted binding sites which correspond to $(x_p(i), y_p(i), z_p(i))$ coordinates to the Cartesian system of the crystal $(x_c(i), y_c(i), z_c(i))$. The coordinate s $(x'(i), y'(i), z'(i))$ correspond to the transformed system following the rotation and translation.

Formulation (111) constitutes a special case of global optimization problems since it involves the minimization of a convex function subject to a set of linear equality and nonconvex equality constraints $RR^\top = I$. The deterministic global optimization algorithm αBB [5, 9, 3, 4, 2], presented briefly in previous sections, is used for the solution of this global optimization problem. The results obtained for the superposition of the predicted HLA-DR3 and I-E^k binding sites with the crystallographic data are presented in the following sections. Four tests are performed in order to evaluate the prediction accuracy of our methodology.

(i) for each predicted binding site the root-mean-square-deviations of Cartesian coordinates of all the atoms (cRMSD) and the C^α atoms are evaluated,

(ii) for each one of the substituted residues, the cRMSD is evaluated considering all the atoms,

(iii) for each one of the substituted residues, a relative-cRMSD is evaluated based on the following formula:

$$
R - cRMSD = \frac{1}{N_{atom}} \sqrt{\sum_{i \in N_{atom}} \frac{1}{3} [\frac{(x_p(i) - x_c(i))}{x_c(i)}^2 + \frac{(y_p(i) - y_c(i))}{y_c(i)}^2 + \frac{(z_p(i) - z_c(i))}{z_c(i)}^2]}
$$

to measure the relative predictive error of the procedure,

(iv) computational binding studies are performed to compare the energetic-based rank ordering of the amino acids in the predicted binding site versus the rank ordering of the amino acids in the binding site based on the crystallographic data, as discussed in later sections.

Prediction of HLA-DR3 Binding Sites We applied our approach to the prediction of the three dimensional structure of HLA-DR3 binding sites.

As presented in Table 37, by substituting Gly to Val in position β86 in pocket 1 of HLA-DR1, the pocket 1 of HLA-DR3 is formulated. The predicted pocket of HLA-DR3 is shown in Figure 55 with the crystallographically obtained pocket superposition. The cRMSD difference between these two pockets is found to be 1.09$\mathring{A}$ based on the differences of the coordinates of all the atoms involved in the pocket. The relative cRMSD for the whole binding site is 0.0425 which corresponds to 4.25% difference of the predicted Cartesian coordinates of the binding site and the crystallographic data. The cRMSD difference based on the α carbons is 0.55$\mathring{A}$. The cRMSD for the substituted residue (Val) is 1.584 $\mathring{A}$ and the relative-cRMSD is 0.04601 which indicates a 4.6% difference between the

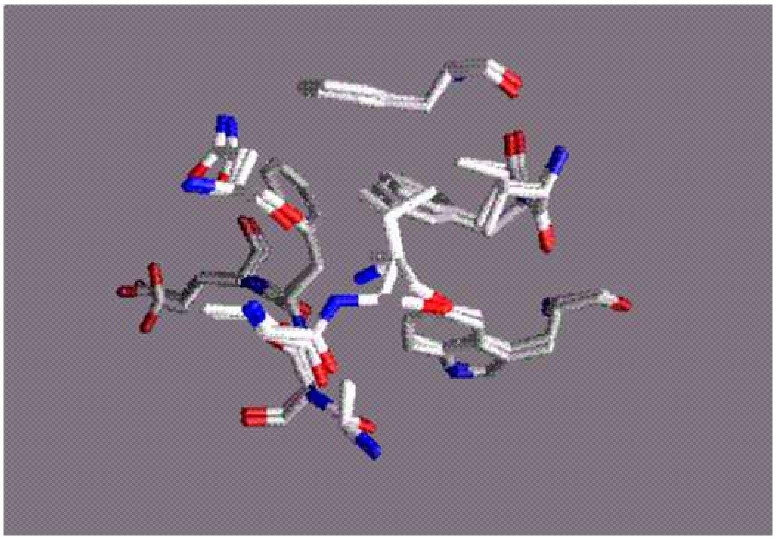

Figure 55: Superposition of the Predicted Pocket 1 of HLA-DR3 vs. Crystallographic Data

predicted valine and the valine determined based on the crystallographic data of the HLA-DR3 molecule [78].

To generate pocket 4 of HLA-DR3, three substitutions are made on the composition of the pockets of HLA-DR1 at the positions β13: Phe $\rightarrow$ Ser, β26: Leu $\rightarrow$ Tyr, and β74: Ala $\rightarrow$ Arg. The cRMSD difference for all the residues in the pocket is 1.11$\AA$, and the overall relative difference of the predicted pocket compared to the crystallographic data is 2.08%. The cRMSD difference based on the α carbons is 0.49 $\AA$. The cRMSD for each one of the substituted residues is 1.67$\AA$ for Ser, 0.83$\AA$ for Tyr and 1.46 $\AA$ for Arg and correspond to relative differences of 3.2%, 1.2% and 2.3%, respectively.

For pocket 6 of HLA-DR3, the substitutions are at positions β11: Leu to Ser, β13: Phe to Ser, and β71: Arg to Lys. The cRMSD difference for this pocket is 1.22$\AA$ based on all atom deviations, which corresponds to a relative-cRMSD of 4.9%. The cRMSD difference based on the α carbons is 0.61 $\AA$. The individual cRMSD for Serβ11 is 1.26 $\AA$, for Serβ13 is 1.62 $\AA$, and for Lys β71 1.82$\AA$ which correspond to relative predictive errors of 7.4%, 3.7% and 3.2%, respectively.

For pocket 7 of HLA-DR3 three substitutions are made at the positions β28: Glu to Asp, β47: Tyr to Phe, and β71: Arg to Lys. The cRMSD difference for this pocket is 1.94$\AA$ based on all atom deviations, which corresponds to a 4.69% deviation. The cRMSD difference based on the α carbons is 0.71 $\AA$. The cRMSD for each one of the substituted residues are 1.08$\AA$ for Phe, 3.08$\AA$ for Asp and 3.4 $\AA$ for Arg and correspond to relative differences of 1.4%, 5.1% and 4.7%, respectively.

Finally, for pocket 9 only one substitution is needed, namely Trp to Glu in position β9 to obtain pocket 9 of HLA-DR3 from pocket 9 of HLA-DR1. The resulting pocket is found to have a cRMSD difference of 1.03 $\AA$ based on all atoms and 0.56$\AA$ based on the C^{α} atoms. The relative-cRMSD based on all atom deviations is 37.2%. Considering only the substituted residue, the cRMSD is 1.67 $\AA$. The large predictive deviation in this pocket is due to the large inherent deviation between the

Table 38: Results for HLA-DR3 Prediction

Pocket	Pocket		Substituted Residues	
	cRMSD ($\mathring{A}$)		Relative-cRMSD (%)	cRMSD ($\mathring{A}$)
	All atoms	C^{α}	All atoms	
1	1.09	0.55	4.6	Val: 1.58
4	1.11	0.49	2.1	Ser: 1.67 Tyr: 0.83 Arg: 1.46
6	1.22	0.61	4.9	Ser: 1.26 Ser: 1.62 Lys: 1.82
7	1.94	0.71	4.7	Asp: 3.08 Phe: 1.08 Lys: 3.40
9	1.32	0.56	37.2	Glu: 1.67

HLA-DR1 and the HLA-DR3 crystallographic data. This cRMSD difference for pocket 9 is 1.05$\mathring{A}$, which corresponds to an inherent relative cRMSD of 20.7%.

The results of our prediction approach for all the pockets are summarized in Table 38. Note that the percentage predictive error is less than 5%, except for pocket 9 where the large inherent deviation between the two crystals prohibits a more accurate prediction.

The coordinates of N and C' are variables in this formulation with bounded ranges for their values around the corresponding atoms in HLA-DR1. Based on the differences observed in the N and C' (x,y,z) coordinates of the HLA-DR1, HLA-DR3, and I-E^k crystals [221, 78, 73] after superposition, tight bounds in the range of [0.3-1.0] suffice. To study further the effect of the bounds we considered bound variations of $(\pm 0.5), (\pm 0.7)$ and (± 1.0). The predicted structures of pocket 1 exhibited small cRMSD differences of 1.18, 1.11 and 1.09 $\mathring{A}$, respectively, calculated based on all atoms.

Our prediction approach considers the simultaneous substitution of all amino acids responsible for the differences of MHC class II molecules. The required substitutions usually involve 2, 3, or 4 residues and give rise to a global optimization problem that includes as variables the dihedral angles of each residue as well as the translation vector and Euler angles defining the relative position of each residue. In order to reduce the size of the resulting global optimization problem, the following two simplifying alternative procedures also were explored. The first approach is sequential in nature. Instead of considering all amino acids substitutions simultaneously, we considered them sequentially. In particular, the conformation of the first considered substituted amino acid was determined by minimizing the intra- and inter-molecular interactions between the specific amino acid and the other residues of the HLA-DR1 binding site. Then, this residue was considered as part of the pocket and the structure of the second substituted residue was determined. In the

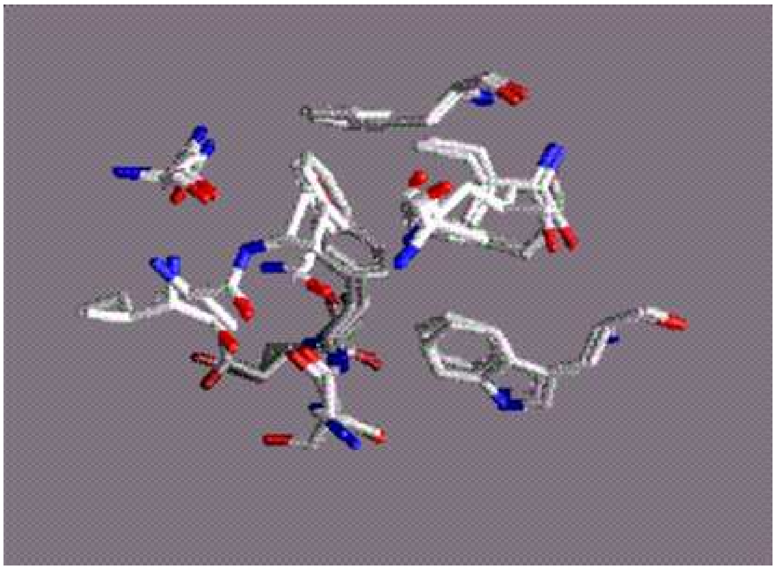

Figure 56: Superposition of the Predicted Pocket 1 of I-E^k vs. Crystallographic Data

second alternative approach we considered each of the substituted amino acids independently and determined their conformations based on minimized energy interactions with the rest of amino acids involved in the pocket of HLA-DR1 molecule. The results obtained for the case of pocket 1 of HLA-DR3 are better than that of the sequential approach having an cRMSD of 2.17 Å compared to 2.51 Å of the sequential procedure but worse than that of the simultaneous approach (cRMSD=1.09Å). The reason is that in the sequential approach the error from the first determined amino acid conformation is accumulated as its conformation affects greatly the conformation of the other sequentially considered amino acids.

Prediction of I-E^k Binding Sites Pocket 1 of I-E^k, requires three substitutions: β85: Val $\rightarrow$ Ile, β86: Gly $\rightarrow$ Phe, and β90: Thr $\rightarrow$ Leu. The predicted pocket is illustrated in Figure 56 with the crystallographic data of I-E^k [78]. The cRMSD difference based on all atoms deviations is 1.67 and corresponds to 9.2% relative predictive error. The cRMSD differences for the individual substituted residues are 2.45, 3.36, and 1.76 Å, for Ile, Phe and Leu, respectively.

For pocket 4 of I-E^k there are four substitutions needed, as shown in Table 37 (β13: Phe to Ser, β74: Ala to Glu, β78: Tyr to Val, and β:71 Arg to Lys). The cRMSD difference is 1.58Å, which corresponds to 3.49% predictive error. For the individual substituted residues the cRMSD differences are 0.78, 1.35, 2.88, and 1.61 Å, for Ser, Glu, Val, and Lys, respectively. These individual differences correspond to relative predictive errors of 1.59%, 2.16%, 4.48%, and 2.03%.

For pocket 6 of I-E^k three substitutions are required at the positions β11: Leu $\rightarrow$ Ser, β13: Phe $\rightarrow$ Cys, and β71: Arg $\rightarrow$ Lys. The cRMSD difference is 1.28 Å based on all atoms, which corresponds to 5.19% relative predictive error. For the individual substituted residues, the cRMSD differences are 1.89, 2.67, and 1.64 for Ser, Cys, and Lys, respectively. These differences correspond

Table 39: Results for I-E^k Prediction

Pocket	Pocket		Substituted Residues	
	cRMSD ($\mathring{A}$)		Relative-cRMSD (%)	cRMSD ($\mathring{A}$)
	All atoms	C^α	All atoms	
1	1.67	0.47	9.2	Ile: 2.45 Phe: 3.36 Leu: 1.76
4	1.58	0.83	3.5	Ser: 0.78 Glu: 1.35 Val: 2.88 Lys: 1.61
6	1.28	0.65	5.2	Ser: 1.89 Cys: 2.67 Lys: 1.64
7	2.03	0.93	4.3	Val: 2.89 Phe: 2.15 Phe: 2.20 Lys: 3.23
9	1.35	0.63	23.3	Val: 1.56 Glu: 2.46 Asn: 1.56

to 4.41%, 14.06% and 2.82% relative predictive error.

Pocket 7 of I-E^k requires four substitutions, as shown in Table 37 (β28: Glu to Val, β47: Tyr to Phe, β67: Leu to Phe, and β71: Arg to Lys). The cRMSD difference is 2.03 $\mathring{A}$ and corresponds to 4.33% relative predictive deviation. For the individual residues the cRMSD differences are 2.89, 2.15, 2.20 and 3.23 $\mathring{A}$ for Val, Pheβ47, Pheβ67, and Lys, respectively, and correspond to 3.95%, 3.1%, 5.28% and 4.41% relative predictive deviation.

Finally, pocket 9 of I-E^k features three substitutions: α72 Ile to Val, β9: Trp to Glu, and β60: Tyr to Asn. The cRMSD difference is 1.35 $\mathring{A}$ which corresponds to 23.3% relative predictive deviation. For the individual residues the cRMSD differences are 1.56, 2.46, and 1.56 $\mathring{A}$ for Val, Glu, and Asn, respectively. The larger relative predictive deviation for this pocket is mainly due to the large relative error for Val at position α72, and the large deviation between the crystals HLA-DR1 and HLA-DR3 which gives a cRMSD of 1.09$\mathring{A}$ and a 21.4% relative deviation. The results for all the pockets are summarized in Table 39.

In order to study the effect of considering different bounds on N and C' coordinates, the proposed approach was applied to all the pockets for ±0.5 and ±0.3 $\mathring{A}$ bounds around the coordinates of the corresponding atoms of HLA-DR1 molecule. The results are shown in Table 40. As was found from the crystallographic data of the I-E^k molecule binding with different peptides (i.e., a peptide derived from murine hemoglobin Hb(64-76), or a peptide from murine heat shock protein

Table 40: Effect of Different Bounds on N and C' Coordinates (I-E^k)

Pocket	**Bounds**	cRMSD ($\mathring{A}$)
1	$\pm$ 0.5	2.26
	$\pm$ 0.3	1.67
4	$\pm$ 0.5	1.81
	$\pm$ 0.3	1.58
6	$\pm$ 0.5	1.28
	$\pm$ 0.3	1.44
7	$\pm$ 0.5	3.17
	$\pm$ 0.3	2.41
9	$\pm$ 0.5	1.84
	$\pm$ 0.3	1.77

Table 41: cRMSD Differences Between HLA-DR1, HLA-DR3 and I-E^k Crystals

Pocket	HLA-DR1 vs HLA-DR3 Crystals - All atoms cRMSD ($\mathring{A}$)	HLA-DR1 vs I-E^k-HB Crystals - All atoms cRMSD ($\mathring{A}$)
1	1.03	1.24
4	0.84	1.23
6	0.84	0.84
7	0.996	0.997
9	1.05	1.092

70 Hsp(236-248)), there is some inherent variability in the range of 0.01-0.4 $\mathring{A}$ for N and C' coordinates. These differences correspond to pocket flexibility to accommodate different peptides.

The obtained cRMSD data for all predicted pockets show good agreement with the crystallographic data considering that there is an inherent difference between the crystals, as shown in Table 41. The cRMSD differences shown in Table 41 represent the differences in the common atoms of the pockets of HLA-DR1 and HLA-DR3 crystals, as well as the differences between HLA-DR1 and I-E^k crystals. These cRMSD differences serve as a reference point in the evaluation of the predicted pockets. For instance, for pocket 1 of HLA-DR3, the predicted structure via the proposed approach has a cRMSD difference of 1.09 $\mathring{A}$ while the crystallographic data of pocket 1 for HLA-DR1 and pocket 1 of HLA-DR3 exhibit a cRMSD of 1.03 $\mathring{A}$ among their common atoms. Comparing the results shown in Tables 38, 39, and 41, it is evident that the predicted structures are close to their reference points, even for pocket 9.

Our approach couples the modeling of energetic interactions and deterministic global optimization approaches and can predict the pockets of HLA-DR3 and I-E^k with small (rms) differences.

C Prediction of Relative Binding Affinities

We have developed a theoretical approach that determines the relative binding affinities of amino acids binding to the five pockets of the MHC II molecule HLA-DR1. MHC II molecules such as HLA-DR1 are cell surface glycoproteins that play a pivotal role in the development of an effective immune response. An important function of HLA molecules is to bind and present antigen peptides to T-cells. Presently there is no comprehensive way of predicting and energetically evaluating peptide binding for HLA molecules.

To determine quantitatively the relative binding affinities of different peptides for HLA molecules, we developed a decomposition approach based on deterministic global optimization that takes advantage of the topography of the HLA binding grove. Our computational results for binding the twenty naturally occurring amino acids in the five pockets of the HLA allele HLA-DRB1*0101 are in excellent agreement with experimental binding assays and with x-ray crystallography data.

1 Definition of Problem

Class II histocompatibility molecules are cell surface molecules that form complexes with self and non-self peptides and then present them to T-cells as part of the immune response. MHC II molecules are important in autoimmune diseases such as diabetes and rheumatoid arthritis, and being able to predict and design the sequences and affinities of peptides which bind to MHC II molecules would increase our understanding of these dieseases as well as our ability to design drugs to treat them.

The question that is addressed is stated as follows:

> *"Given the (x,y,z) coordinates of the atoms in HLA-DR1 [221], can we predict the affinity and conformation of the peptides which bind to it?"*

2 Approach

We have developed a decomposition approach for predicting the binding affinity and conformation of peptides binding to HLA-DR1. Our decomposition approach takes advantage of the fact that the binding affinity of a peptide for HLA-DR1 molecules is determined primarily by the binding affinity of individual amino acid residues for HLA-DR1's five binding pockets. Our approach uses a sequence of three steps [12]:

I: Consideration of each binding pocket individually

II: Evaluation of the binding of one amino acid at a time to a given pocket

III: Creation of a rank-ordered list of strong, average, and weak amino acid binders for each pocket

Step I involves determining which residues of the HLA-DR1 molecule compose a given pocket. This process is discussed in Section V.C.3 below. Step II involves formulating a mathematical model for the potential and solvation energy of the pocket and the binding amino acid, and then using this model to predict the amino acid conformation which corresponds to the global minimum potential and solvation energy state of the system. This global minimum energy state is considered to be the system's most stable state. The mathematical model used to describe the energy of the HLA-DR1/peptide system is discussed in Section V.C.3 below, while the global minimization algorithm used to find the peptide conformation corresponding to the global minimum energy is discussed in Section V.C.4 below. Step III involves comparison of the amino acids binding to a given pocket. The comparison standard used is the change in potential and solvation energy of an amino acid on binding, ΔE. This quantity is defined as the difference between the global minimum potential and solvation energy of an amino acid when it is bound in the pocket (E_{Total}) and the global minimum potential and solvation energy of a free amino acid far from the pocket or any other interactions (E^o_{Res}):

$$\Delta E \;=\; E_{Total} - E^o_{Res} \tag{112}$$

The quantity ΔE can be thought of as the difference between the final (bound) and initial (free) states of an amino acid. Thermodynamics predicts that events will proceed in the direction which lowers the total energy of their components. Thus ΔE is a measure of the tendency of an amino acid to bind with the pocket of the HLA-DR1 molecule. A very negative ΔE corresponds to very strong binding.

3 Modeling

Pocket Definition Consideration of each of HLA-DR1's five binding pockets independently, which corresponds to Step I in Section V.C.2 above, involves determining which residues of the HLA-DR1 molecule compose a given pocket. X-ray crystallography data are available that provide the (x,y,z) Cartesian coordinates of the atoms in the complex of HLA-DRB1*0101 and the influenza peptide HA [221]. The Program for Pocket Definition (PPD) is able to define a given HLA-DR1 pocket from this crystallographic data by calculating which HLA-DR1 amino acids have atoms within a radius $\mathcal{R}$ of the atoms of the influenza peptide amino acid bound to the pocket [12]. The HLA-DR1 residues which do have atoms within this radius constitute the pocket. The inputs required for PPD operation are the value of $\mathcal{R}$, the crystallographic data for the entire HLA-DR1/peptide complex [221], and the crystallographic data for the peptide amino acid bound in the HLA-DR1 pocket. The crystallographic coordinates of the pocket residues are given in an output file. A range of $\mathcal{R}$ values has been evaluated [12] in order to determine an appropriate radius which represents a pocket realistically but which does not include so many residues in the pocket that energy minimization is computationally intractable. Table 42 presents the residues defining each of HLA-DR1's five pockets at different radii. The general trends of this table include increased pocket complexity with increased radius (such as in Pocket 1), constant pocket complexity despite increased radius (such as in Pocket 7), and the much larger number of amino acids in Pocket 1 in comparison to the other four pockets [12]. Based on the results in Table 42, a radius of 5 Å was used to define Pockets 1, 4, 6, and 7 of HLA-DR1, while a radius of 4.5 Å was used to define Pocket 9.

Table 42: PPD Pocket Compositions for $\mathcal{R}$ = 4.0-5.0 $\AA$

Pocket	$\mathcal{R} =$	**4.0**	$\mathcal{R} =$	**4.5**	$\mathcal{R} =$	**5.0**
1	ileα31	pheα32	ileα31	pheα32	ileα31	pheα32
	trpα43	alaα52	trpα43	alaα52	trpα43	alaα52
	serα53	pheα54	serα53	pheα54	serα53	pheα54
	valβ85	glyβ86	valβ85	glyβ86	valβ85	glyβ86
	pheβ89		pheβ89	pheα24	pheβ89	pheα24
			asnβ82		asnβ82	gluα55
					thrβ90	
4	glnα09	asnα62	glnα09	asnα62	glnα09	asnα62
	pheβ13	glnβ70	pheβ13	glnβ70	pheβ13	glnβ70
	argβ71	alaβ74	argβ71	alaβ74	argβ71	alaβ74
	tyrβ78		tyrβ78	gluα11	tyrβ78	gluα11
			leuβ26		leuβ26	
6	gluα11	asnα62	gluα11	asnα62	gluα11	asnα62
	valα65	aspα66	valα65	aspα66	valα65	aspα66
	leuβ11		leuβ11		leuβ11	pheβ13
					argβ71	
7	valα65	asnα69	valα65	asnα69	valα65	asnα69
	gluβ28	tyrβ47	gluβ28	tyrβ47	gluβ28	tyrβ47
	trpβ61	leuβ67	trpβ61	leuβ67	trpβ61	leuβ67
	argβ71		argβ71		argβ71	
9	ileα72	asnα69	ileα72	asnα69	ileα72	asnα69
	metα73	argα76	metα73	argα76	metα73	argα76
	trpβ09	aspβ57	trpβ09	aspβ57	trpβ09	aspβ57
	tyrβ60		tyrβ60	trpβ61	tyrβ60	trpβ61
					leuα70	

Problem Formulation The position of a particular peptide or amino acid chain in space can be described completely by a translation vector, a rotation matrix, and a set of dihedral angles. The translation vector is defined as the coordinates of the backbone nitrogen atom on the first residue of a chain. The rotation matrix is defined by the Euler angles of the first chain residue. In our work, the HLA-DR1 pockets are considered rigid and fixed. Thus the variables are the nitrogen coordinates, Euler angles, and dihedral angles of the amino acid which is attempting to bind to a pocket.

Because the decomposition approach described in Section V.C.2 above implicitly assumes that the binding residue is part of a longer peptide, the Cartesian coordinates of the carboxyl carbon atom (C') must be constrained. The decomposition approach is based on the assumption that the rest of the peptide, although not explicitly modeled, is binding normally, and thus that the backbone atoms of the binding peptide do not vary greatly from their crystallographic positions. Because the optimization problem is formulated on internal coordinates, the Cartesian coordinates of C' are implicit variables defined as functions of the translation vector, Euler angles, and dihedral angles of the peptide [12].

With these variables in mind, formulation of the energy minimization problem proceeds as follows [12]. Let E be the function which calculates the potential and solvation energy of the HLA-DR1 pocket/binder system. Let the Cartesian coordinates of the nitrogen translation vector be defined by the variables N_x, N_y, and N_z. Let the Euler angles be represented by ε_1, ε_2, and ε_3. Let $k = 1, \ldots, K$, where K is the total number of side chain dihedral angles of the amino acid residue binding to a pocket. The set of variable dihedral angles then includes the backbone dihedral angles ϕ, ψ, and ω, and the side chain angles χ_k. The Cartesian coordinates of the backbone carboxyl carbon (C'), are defined by C'_x, C'_y, and C'_z. Utilizing these variable definitions, the potential energy minimization problem can be formulated as follows:

$$\min \quad E\left(\phi, \psi, \omega, \chi^k, N_x, N_y, N_z, \varepsilon_1, \varepsilon_2, \varepsilon_3\right) \tag{113}$$

$$\text{s.t.} \quad -\pi \leq \phi \leq \pi \tag{114}$$

$$-\pi \leq \psi \leq \pi \tag{115}$$

$$-\pi \leq \omega \leq \pi \tag{116}$$

$$-\pi \leq \chi^k \leq \pi, \quad k = 1, \ldots, K \tag{117}$$

$$-\pi \leq \varepsilon_1 \leq \pi \tag{118}$$

$$-\pi \leq \varepsilon_2 \leq \pi \tag{119}$$

$$-\pi \leq \varepsilon_3 \leq \pi \tag{120}$$

$$N_x^l \leq N_x \leq N_x^u \tag{121}$$

$$N_y^l \leq N_y \leq N_y^u \tag{122}$$

$$N_z^l \leq N_z \leq N_z^u \tag{123}$$

$$C_x'^l \leq C_x'\left(\phi,\psi,\omega,\chi^k,N_x,N_y,N_z,\varepsilon_1,\varepsilon_2,\varepsilon_3\right) \leq C_x'^u \tag{124}$$

$$C_y'^l \leq C_y'\left(\phi,\psi,\omega,\chi^k,N_x,N_y,N_z,\varepsilon_1,\varepsilon_2,\varepsilon_3\right) \leq C_y'^u \tag{125}$$

$$C_z'^l \leq C_z'\left(\phi,\psi,\omega,\chi^k,N_x,N_y,N_z,\varepsilon_1,\varepsilon_2,\varepsilon_3\right) \leq C_z'^u \tag{126}$$

In the formulation above, the superscripts u and l denote upper and lower bounds, respectively, for the Cartesian coordinates of both the amino nitrogen and the carboxyl carbon.

Although the constraints on the amino nitrogen in the formulation above can be considered directly as problem variables, the C' coordinates are not explicit variables and consequently must be defined as a function of the other variables [12]. Because the energy minimization problem described above involves these implicit constraints on the location of C', a penalty function must be added to the function E in order to implement these constraints. The modified form of the function E is then [12]:

$$E' = E + \beta\{ \quad \langle C_x'^l - C_x'\rangle + \langle C_x' - C_x'^u\rangle + \tag{127}$$

$$\langle C_y'^l - C_y'\rangle + \langle C_y' - C_y'^u\rangle + \tag{128}$$

$$\langle C_z'^l - C_z'\rangle + \langle C_z' - C_z'^u\rangle \; \} \tag{129}$$

The $\langle\rangle$ function has the following definition: $\langle\mathcal{A}\rangle$ equals $\mathcal{A}$ if $\mathcal{A}$ is greater than zero, otherwise $\langle\mathcal{A}\rangle$ equals zero. Therefore, if the coordinates of C' are within their respective bounds, the function E will not be modified. If, however, a particular coordinate falls outside of its bounds, the function will be increased by an arbitrarily large constant β. The conformation's energy then would be arbitrarily large, and the conformation would be discarded as a choice for the minimum energy conformation.

Note that E in the formulation above is a nonconvex function involving numerous local minima that correspond to metastable states of the specific amino acid binding to the pocket 1. A single global minimum defines the energetically most favorable peptide conformation. In establishing a ranked-order list of binding peptides, one needs to identify rigorously the best conformation of (i) the binding residue far from the pocket and (ii) the complex of Pocket 1 with the binding residue. Consequently, there is a need for a method that can guarantee convergence to the global minimum potential energy conformation and which is capable of solving large-scale constrained optimization problems. The global optimization approach αBB [9, 11, 158] is one such method.

GLO-DOCK The αBB algorithm is interfaced and supported with several other programs in the overall energy minimization scheme, and the entire collection of programs is known as GLO-DOCK. The additional programs include MSEED, RRIGS, NPSOL, and PACK. The MSEED program is discussed in Section III.A.2 above and calculates solvent-accessible surface areas, and the RRIGS program is discussed in Section III.A.2 above and calculates solvent-accessible volumes. Only one of these programs is utilized for calculating solvent energies during a given peptide docking optimization. The program NPSOL [79] is a nonlinear local optimization solver used in the calculation of upper and lower bounds for αBB. The PACK program [205], and its associated program prePACK, is a peptide calculation program. The prePACK program initializes PACK by converting the amino acid residue data supplied by the program's user into the format required by the ECEPP/3 potential energy model. The prePACK program also generates the parameter values used by PACK in calculating energy potentials. The PACK program transforms Cartesian coordinates into internal (dihedral angle) coordinates and uses the ECEPP/3 potential energy model to provide function and gradient evaluations to αBB. The PACK program is able to keep track of data for several peptides and make appropriate calls to ECEPP/3 for calculation of their interaction energies [205]. As discussed below, solvation contributions based on solvent-accessible area are added only at local minima, so the program MSEED is called from αBB through PACK once a local minimum has been found. The program RRIGS is called from αBB though PACK at each local minimization step.

Several supporting programs generate the input files used in this overall minimization scheme. These programs include PPD and ARAS. PPD, the Program for Pocket Definition, was discussed above and defines a given pocket from crystallographic data. The output file from PPD is then used as an input file for the program ARAS, the Amino acid Residue Angle Solver. ARAS converts the crystallographic data from the PPD output file into translation vectors, Euler angles, and dihedral angles for each amino acid in the file. An ARAS output file (*name.input* in Figure 57) is then used as an input file for PACK. Three other input files are required for peptide docking optimizations: *name.abb* and *prep.name.abb*, which provide the bounds on the initial nitrogen atom and other information needed by αBB; and *name.bounds*, which provides the bounds on C' for the penalty function. The (x,y,z) nitrogen and C' bounds for each pocket binder are determined by examining the crystallographic data for the corresponding peptides in the HLA-DRB1*0101/influenza virus peptide complex presented by Stern et al. [221]. These bounds are set at ± 0.7 Å from the crystallographic coordinates. A schematic diagram for the overall global optimization scheme is given in Figure 57.

Solvation Methods Because the polar, cohesive nature of water profoundly affects all molecular interactions in biological systems [224], the effects of solvation on the conformation of a protein must be included in an accurate protein model.

There are many types of solvation models. Explicit solvation models arrange individual water molecules around peptides and calculate solvent/peptide interactions with potential models similar to those discussed in Section III.A.1 above. These explicit models are prohibitively expensive computationally because of the large number of water molecules involved and because a given peptide conformation has a large number of equivalent possible water molecule arrangements, making it necessary to calculate the energy of many solvent arrangements and average them together [128]. Neglecting the molecular nature of water molecules yields much simpler, implicit solvation models. Models of this type often estimate energies of solvation as functions of solvent-accessible surface areas or volumes.

Our work is based on two separate implicit methods of determining solvation potentials. One method involves solvent-accessible area calculations and the other involves solvent-accessible volume calculations. These models are based on two assumptions: that a solvation energy can be calculated for each functional group of a peptide by calculating an averaged energy of interaction between

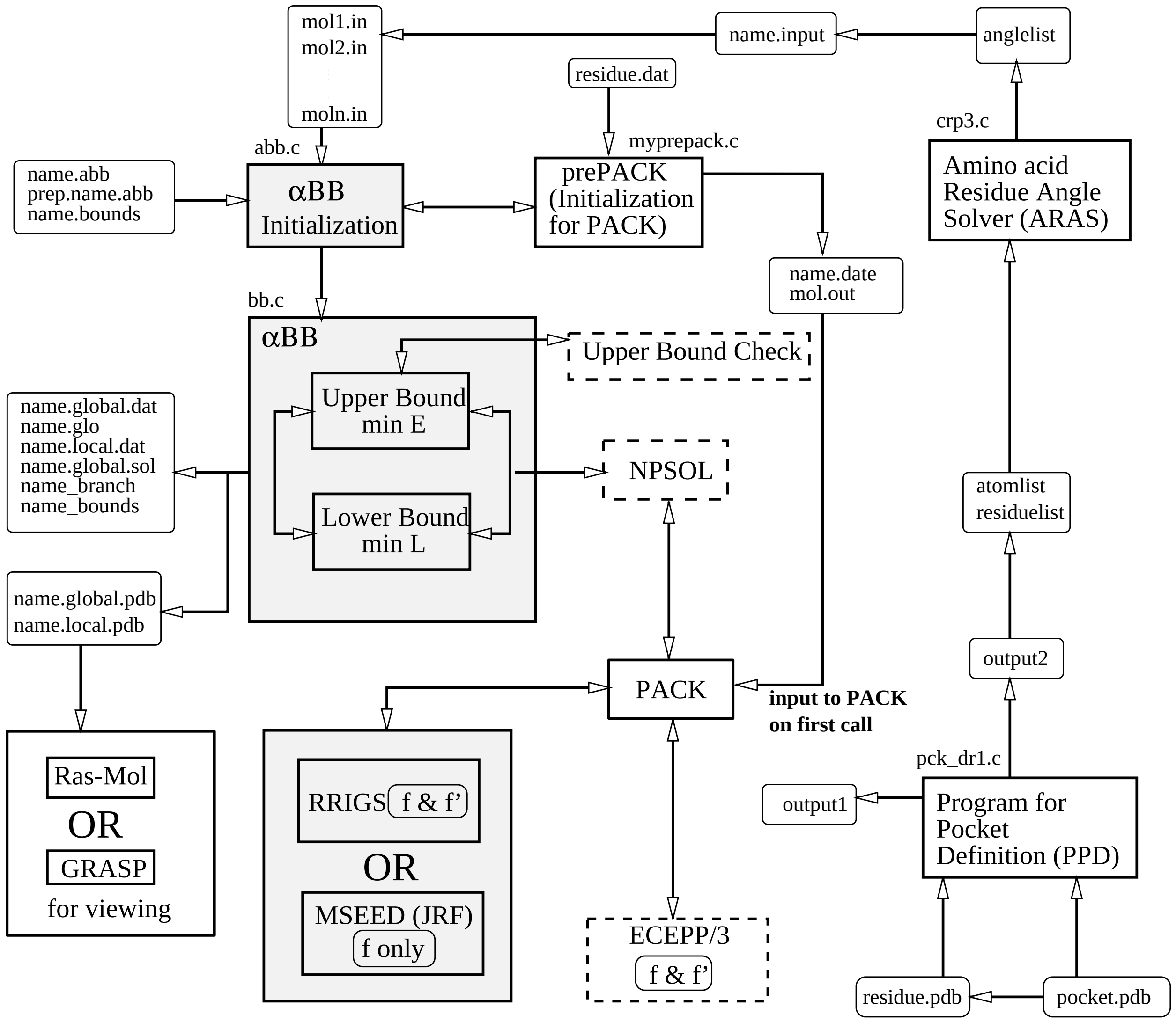

Figure 57: Schematic Diagram for Peptide Docking Global Optimization: The arrows indicate the direction of information flow. The names of input, output, and source code files are indicated. References to "f & f'" and "f only" describe whether gradient evaluations or only function evaluations are used.

the group and a layer of solvent known as the solvation shell, and that these solvation energies are additive. Thus the model assumes that the total energy of solvation of a peptide can be expressed as the sum of the energies of solvation for each of the functional groups of the peptide [128].

The solvent-accessible area solvation model used in our work is based on a program called MSEED [191]. This model assumes that the energy of solvation is proportional to the solvent-accessible surface area of the peptide, as discussed in Section III.A.2 above. MSEED area calculations have some limitations, however. First, MSEED does not always search effectively for the peptide's surface areas. The error incurred by this ineffective search, however, has been shown to be less than 2% for a number of test problems [191]. Second, changes in peptide conformations produced by minimization of the total energy of the peptide proceed continuously but not necessarily smoothly, and surface area gradients may thus have discontinuities. Large discontinuities may cause minimization techniques which require calculation of first derivatives to fail to converge. This problem is avoided in our work because gradients for area-accessible solvation contributions are not calculated and surface-accessible solvation energies are included in the total energy only at local minimum energy conformations and are not part of local minimization processes.

The solvent-accessible volume model used in our work is based on a program called RRIGS, which stands for Reduced Radius Independent Gaussian Sphere [14]. This model assumes that the energy of solvation of a peptide is proportional to the solvent-accessible volume of a solvation layer or shell around the peptide, as discussed in Section III.A.2 above. This method provides continuous derivatives of the solvation potential, so solvation contributions to total energy can be added at every step of local minimizations and not just at the local minimum itself. Thus the RRIGS solvation model interfaces well with the ECEPP/3 potential energy model [128].

Combining the ECEPP/3 potential energy model with a solvation model creates an expression for the total potential and solvation energy (E_{Total}) of the system: $E_{Total} = E_{Potential} + E_{Solvation}$, where $E_{Potential}$ is calculated from ECEPP/3 and $E_{Solvation}$ is calculated from either the MSEED or RIGGS solvation models. With this mathematical model for the potential and solvation energy of the pocket and the binding amino acid in place, the next step in evaluating the binding of one amino acid at a time to a given HLA-DR1 pocket is finding the amino acid conformation which corresponds to the global minimum potential and solvation energy of the system.

4 Deterministic Global Optimization

The first step in implementing a global optimization algorithm like αBB is the formulation of the optimization problem. This involves choosing which functions will be optimized (either minimized or maximized), choosing the variables which will be optimized, and choosing the constraints which will be included in the problem. For the peptide docking prediction problem, implementing a global optimization algorithm also involves deciding whether to minimize the total energy function based on the Cartesian coordinates of the peptide atoms or based on the dihedral angles of the peptide. Because optimization constraints are more easily applied to internal coordinates like dihedral angles than to Cartesian coordinates [158], we used internal coordinates for our work. The problem formulation is developed in Section V.C.3 above. The function E shown in Section V.C.3 is difficult to minimize because it is nonlinear, nonconvex, and has multiple local minima. These local minima correspond to metastable states of the amino acid binder being modeled, but the single global minimum is the minimum which defines the energetically most stable peptide conformation.

Our minimization scheme determines the peptide conformation that corresponds to the global minimum total potential and solvation energy through a series of steps [12, 128]:

1. Upper bound calculation: The local solver NPSOL identifies a local minimum of the potential energy function supplied by PACK in a region (rectangle) defined by the lower and upper bounds of the variables. These bounds are supplied by αBB. If solvent-accessible volume is being considered, potential energy evaluations during local minimization are made using the ECEPP/3 model and RRIGS. If solvent-accessible surface area is being considered, potential energy evaluations are made using only the ECEPP/3 model and the solvation energy is calculated by MSEED and added only at the local minimum.

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

3. The current rectangle (region) is partitioned by bisecting its longest side.

4. Lower bound calculation: The convex underestimator function L is minimized in each new rectangle using NPSOL and PACK (with ECEPP/3). If solvent-accessible volume is being considered, potential energy evaluations are also made using RRIGS. If solvent-accessible surface area is being considered, potential energy evaluations are not made with MSEED, and the solvation contributions are added only at the local minimum. If a minimum is greater than the best upper bound, the corresponding rectangle will be eliminated from the search. Otherwise, the local minimum value is stored.

5. The rectangle with the current minimum value for L is selected for further partitioning.

6. If the best upper and lower bounds are within the user-specified tolerance ϵ the program will finish, otherwise it will proceed back to Step 1.

Repeating this optimization scheme for each naturally occurring amino acid in each of HLA-DR1's five binding pockets and then listing each pocket's amino acid binders in order of increasing global minimum total energy (decreasing binding affinity) creates a rank ordered list of strong, average, and weak amino acid binders for each pocket.

We then introduced an energetic-based criterion to evaluate the energy of interaction between a given pocket and each naturally occurring amino acid. This measure, which we denote as ΔE, corresponds to the difference between (i) the global minimum total potential and solvation energy that considers all the energetic atom-to-atom interactions, classified as inter-interactions between the atoms of the residues that define the pocket of HLA-DR1 protein and the atoms of the considered naturally occurring amino acid, and classified as intra-interactions among the atoms of the considered naturally occurring amino acid, and (ii) the global minimum potential energy of the considered naturally occurring amino acid when it is far away from the pocket. Equation 130 illustrates this criterion:

$$\Delta E \;=\; E^{o}_{Total} - E^{o}_{Res} \tag{130}$$

where E^0_{Total} is the global minimum of the potential energy of the complex of the pocket and the binding peptide or amino acid, and E^o_{Res} is the global minimum of the potential energy of the peptide or amino acid away from the pocket. Note that ΔE does not represent a difference in the free energies of the complex and isolated amino acids. Instead, it denotes the difference between potential and solvation energy for the complex and the isolated amino acids. With this energetic-based criterion, we ranked relative binding affinities.

5 Computational Studies

Binding Affinity Evaluation in HLA-DR1 Pockets The area and volume solvation methods correctly predict the binding affinity and conformations of the strongest binders to Pockets 1, 4, and 6. Neither the area nor the volume solvation methods correctly predict the crystallographic binder conformations for Pocket 7 and 9, perhaps due the the pockets' incomplete definition.

The volume solvation method appears to be a stronger method for considering solvation than the area solvation method. The area solvation method does not use separate parameters for charged and neutral atoms, and its structure does not permit consideration of area contributions at each step of local minimizations of the total energy. Solvation energy contributions in the volume method are of the same order of magnitude as nonbonded contributions, while solvation energy contributions are an order of magnitude larger than nonbonded energy contributions in the area solvation method. The domination of total energy values by solvation in the area solvation method may not distinguish amino acid binders sufficiently from one another in rank-ordered binding lists.

Our global optimization results are in excellent agreement with available experimental data. Experimental data [12] for amino acids binding to Pocket 1 are shown in Figure 58. We were able to reproduce the relative binding affinities shown in the figure, and all of our other relative binding affinity results agree with literature data. The results for Pocket 1 and Pocket 4 are especially encouraging, because Pocket 1 is considered to be the most discriminating and most important pocket for successful peptide binding [92] and Pocket 4 is considered to be one of the most important pockets in T cell recognition interactions [75]. This agreement indicates that our approach is an accurate, effective tool for approaching the peptide docking problem.

The need for determining the conformation of a binding amino acid which corresponds to its global minimum total energy instead of to a local minimum total energy is illustrated in Figure 59. Figure 59 shows a local minimum conformation and the global minimum conformation of tyrosine in Pocket 1 for the volume solvation method. The volume solvation method's local minimum conformation of tyrosine has a ΔE of -17.349 kcal/mol and is shown in a lighter shade, while the global minimum conformation of tyrosine has a ΔE of -20.155 and is shown in darker shade. There is only a 13.9% difference between these two ΔE values, but there is a significant difference in their conformations. The global minimum conformation corresponds closely to the crystallographically determined conformation, highlighting the necessity of not mistaking a local solution for the global solution. This comparison also highlights the need for global optimization methods in approaches to the peptide docking prediction challenge.

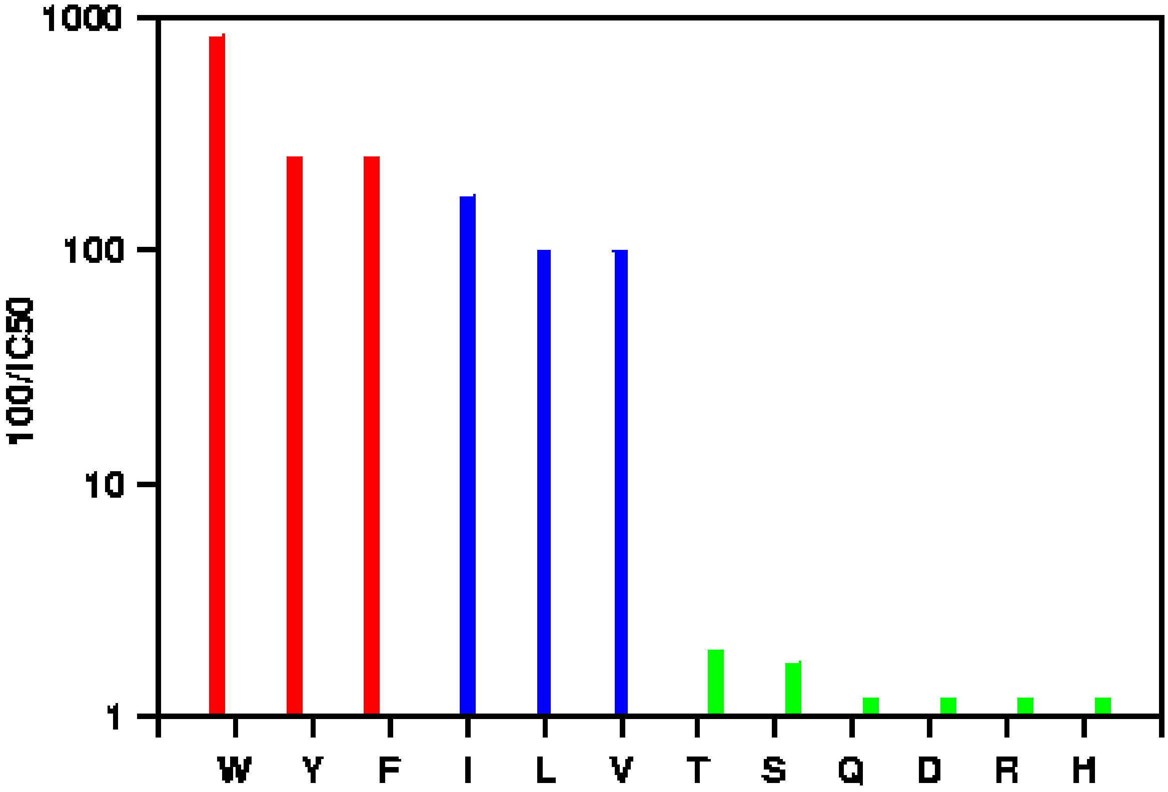

Figure 58: Pocket 1 Competitive Binding Assays

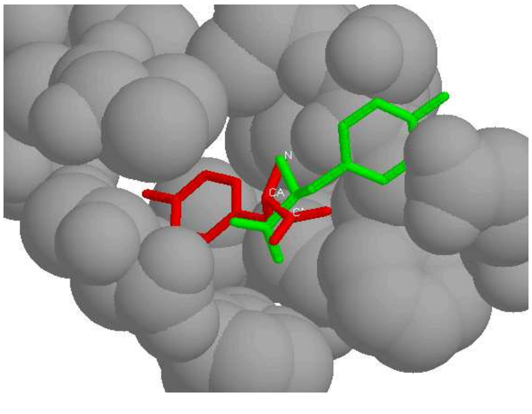

Figure 59: Global (darker) vs. Local (lighter) Tyrosine Conformations in Pocket 1, Volume Solvation

Binding Affinity Evaluation After Structure Prediction Our prediction of the structures of MHC class II binding sites has significant implications for the evaluation of peptide binding to HLA molecules. We applied our biniding affinity prediction methodology to the binding pocket structures we predicted in Section V.B.5. We then compared the results of predicting binding affinities for predicted versus crystallographic pockets.

We applied our methodology to the pocket we predicted for pocket 1 of HLA-DR3 and to the pocket obtained from crystallographic data [221] for the binding amino acids Phe, Ile, and Met. Based on the energy differences we found that Phe is a better binder than Met by 1.1 kcal/mol and that Met is a better binder than Ile by 3.9 kcal/mol for the predicted pocket. For pocket 1 based on the crystallographic data, our binding studies determined the same sequence (i.e., Phe followed by Met and Met by Ile) with corresponding differences of 2.37 kcal/mol for Phe to Met and 2.06 for Met to Ile. Application of our predictive binding approach [12] to the predicted, as well as to the crystallographicaly obtained, pocket 4 of HLA-DR3 for the binding amino acids Asp, Glu, Ile, and Phe showed that the negatively charged Asp and Glu are very strong binders. In contrast, Ile and Phe were weaker binders than Asp and Glu.

We also applied our predictive binding approach to the predicted pocket 1 of I-E^k for the binding amino acids Ile, Val, and Phe. Our results showed that Phe is a better binder than Ile by an energy difference of 6.1 kcal/mol, and that Ile binds better than Val by an energy difference of 2.8 kcal/mol. We obtained similar results from the crystallographic data, with Phe being a better binder than Ile and Ile being a better binder than Val with energy differences of 4.4 and 0.7 kcal/mol, respectively.

In order to verify further the correct prediction of the binding sites of HLA molecules, we used the crystal of HLA-DR3 [78] to predict pocket 1 of HLA-DR1. We compared the results obtained using the predicted pocket to those found with the crystallographically obtained pocket [221]. As shown in Table 43, the binding studies using the predicted pocket illustrate the same trends as the binding studies using the crystallographic pocket. Therefore, our approach not only predicts the binding site structure of class II HLA molecules, but also provides results consistent with the binding studies of individual amino acids based on the crystallographic data.

D Perspectives and Future Work

We currently are expanding and extending our binding site structure and binding affinity prediction methods. We are expanding our methods to incorporate rigorous calculation of free energies. Our approach to these free energy calculations involves the terms in the following equation:

$$E_{Total} \;=\; E_{Vacuum} + TS_{Vacuum} + E_{Cavity} + E_{Solvation} + E_{Ionize} \tag{131}$$

where E_{Total} is the total free energy of a protein-protein system. In this approach, as in our earlier approach discussed above, E_{Vacuum} is the potential energy of a protein system conformation in a vacuum calculated from the ECEPP/3 force field. In our expanded approach, TS_{Vacuum} is the entropy of a protein conformation in a vacuum. In order to calculate this term, we generate a large set of unique conformers and then apply a harmonic approximation to obtain the entropy

Table 43: Comparison of Binidng Studies in Predicted Binindg Sites vs. Crystallographic Binding Sites in Pocket 1 of HLA-DR1 ($\mathcal{R}$ = 5.0 Å), Area Solvation

Residue	**Δ E** Crystal (kcal/mol)	**Δ E** Prediction (kcal/mol)	Difference (kcal/mol)	Difference %
Tyr	-20.000	-18.850	-1.15	5.75
Phe	-19.625	-18.040	-1.58	2.95
Trp	-16.950	-17.754	0.80	4.72
Gln	-15.396	-15.916	0.52	3.37
Met	-13.943	-13.928	-0.02	0.14
Asn	-13.784	-14.644	0.86	6.24
Thr	-13.297	-13.297	0.00	0.00
Leu	-12.481	-12.399	-0.08	0.64
Ile	-12.465	-12.486	0.02	0.16
Ser	-11.557	-11.187	-0.37	3.20
Cys	-11.280	-11.087	-0.19	1.68
Val	-11.209	-11.324	0.12	1.07
Ala	-10.355	-10.338	-0.02	0.19
Gly	-10.091	-9.996	-0.09	0.89
Glu-	-7.744	-6.891	-0.85	10.97
Asp-	-2.431	-2.594	0.16	6.58

of each conformation. The E_{Cavity} term in this approach is the energy required to form a protein conformation's cavity in aqueous solvent. This cavity energy is estimated to be proportional to the surface area of the protein system exposed to water. We calculate the $E_{Solvation}$ term in this expanded approach with Poisson-Boltzmann electrostatics by using the DELPHI software package [80, 81, 209]. The $E_{Solvation}$ term is the difference in a protein system conformation's polarization energy in solvent (dielectric constant ε=80) and in a vacuum (dielectric constant ε=1), as shown in the equation below:

$$E_{Solvation} = \frac{1}{2}\sum_i\sum_s \frac{q_i\sigma_{s,\varepsilon=80}}{|\mathbf{r_i}-\mathbf{r_s}|} - \frac{1}{2}\sum_i\sum_s \frac{q_i\sigma_{s,\varepsilon=1}}{|\mathbf{r_i}-\mathbf{r_s}|} \tag{132}$$

where q_i is the charge associated with atom i and σ_s is the surface charge induced by each charge s other than i. The E_{Ionize} term is the energy due to the ionization state of a protein system at a given pH. These expansions to our binding site structure and binding affinity prediction methods will allow us to model solvent effects more rigorously and more accurately, as well as allowing the study of ionization effects. We then will have the ability to calculate and predict not only relative binding affinities, but accurate, quantitative binding affinitites. The drawback of employing the entropic and Poisson-Boltzmann calculations discussed above is the large increase in computational time they require. We are exploring ways of parallelizing our algorithm to address this issue.

In addition to expanding our methods, we are extending them to the investigation of larger systems. We are examining the role of the peptide residues intermediate to the pocket-binding residues in peptides which bind to HLA molecules, as well as modeling the docking of entire peptides to HLA molecules. Our future plans include extending our methods to the examination of T-cell interactions with HLA molecules and bound peptides.

Our computational and experimental results demonstrate that applying atomistic level modeling and deterministic global optimization is a promising approach to a systematic framework for peptide docking prediction. The strengths of our peptide docking prediction model are its guaranteed convergence to the global minimum energy, its detailed modeling of entropic, electrostatic and other energetic interactions, and its quantitative prediction of binding free energy.

The predictive power of protein-protein interaction and peptide docking models is of significant and increasing importance. Accurate prediction will lead to the more efficient and effective design of drugs and devices. Peptide docking and protein-protein interaction prediction thus will play a valuable role in capitalizing on the data provided by the mapping of human and other genomes.

VI Conclusions

The intense worldwide experimental and theoretical research effort directed toward solving the protein folding and peptide docking problems underscores their importance. The ability to predict computationally the folding of proteins and the formation of protein-protein complexes would support and help direct experimental work in biology, chemistry, biophysics, and pharmaceutical development. In this review we have shown that molecular modeling and global optimization are the dominant factors that will provide solutions to these problems.

In particular, this review has focused on the use of ab-initio models, which give rise to a series of complex mathematical problems. A second important component has been the application of deterministic global optimization, namely the αBB algorithm, for solving the resulting problems. In this review we have analyzed and discussed many issues related to the modeling of protein folding and peptide docking systems. These observations have highlighted the extraordinary difficulty of these problems and the crucial interdependence of ab initio modeling and deterministic global optimization approaches.

Acknowledgements

The authors gratefully acknowledge financial support from the National Science Foundation and the National Institutes of Health (R01 GM52032, 1 F32 GM20007).

References

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

[2] C. S. Adjiman, I. P. Androulakis, and C. A. Floudas. A global optimization method for general twice-differentiable nlps - ii. implementation and computational results. *Comput. Chem. Eng.*, 22:1159–1179, 1998.

[3] C. S. Adjiman, I. P. Androulakis, C. D. Maranas, and C. A. Floudas. A global optimization method, αBB, for process design. *Comput. Chem. Eng.*, 20:S419–S424, 1996.

[4] C. S. Adjiman, S. Dallwig, C. A. Floudas, and A. Neumaier. A global optimization method for general twice-differentiable nlps - i. theoretical advances. *Comput. Chem. Eng.*, 22:1137–1158, 1998.

[5] C. S. Adjiman and C. A. Floudas. Rigorous convex underestimators for general twice–differentiable problems. *J. Glob. Opt.*, 9:23–40, 1996.

[6] F. A. Al-Khayyal and J. E. Falk. Jointly constrained biconvex programming. *Maths Ops Res.*, 8:273–286, 1983.

[7] N. L. Allinger. Conformational analysis. 130. mm2. a hydrocarbon force field utilizing v_1 and v_2 torsional terms. *J. Am. Chem. Soc.*, 99:8127–8134, 1977.

[8] N. L. Allinger, Y. H. Yuh, and J. H. Lii. Molecular mechanics. the mm3 force field for hydrocarbons. *J. Am. Chem. Soc.*, 111(23):8551–8565, 1989.

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

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

[11] I.P. Androulakis, C.D. Maranas, and C.A. Floudas. Prediction of oligopeptide conformations via deterministic global optimization. *Journal of Global Optimization*, 11:1–34, 1997.

[12] I.P. Androulakis, N.N. Nayak, M.G. Ierapetritou, D.S. Monos, and C.A. Floudas. A predictive method for the evaluation of peptide binding in pocket 1 of hla-drb1 via global minimization of energy interactions. *Proteins*, 29:87–102, 1997.

[13] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White Jr. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. USA*, 47:1309–1314, 1961.

[14] J. D. Augspurger and H. A. Scheraga. An efficient, differentiable hydration potential for peptides and proteins. *J. Comp. Chem*, 17:1549–1558, 1996.

[15] D. J. Bacon and J. Moult. Docking by least-squares fitting of molecular surface patterns. *J. Mol. Biol.*, 225:849–858, 1992.

[16] R. L. Baldwin and G. D. Rose. Is protein folding hierarchic? i. local structure and peptide folding. *Trends Biochem. Sci.*, 24(1):26–33, 1999.

[17] R. L. Baldwin and G. D. Rose. Is protein folding hierarchic? ii. folding intermediates and transition states. *Trends Biochem. Sci.*, 24(2):77–83, 1999.

[18] K. D. Ball and R. S. Berry. Realistic master equation modeling of relaxation on complete potential energy surfaces: Partition function and equilibrium results. *J. Chem. Phys.*, 109(19):8541–8556, 1998.

[19] K. D. Ball and R. S. Berry. Realistic master equation modeling of relaxation on complete potential energy surfaces: Kinetic results. *J. Chem. Phys.*, 109(19):8557–8572, 1998.

[20] A. Banerjee, N. Adams, J. Simons, and R. Shepard. Search for stationary points on surface. *J. Phys. Chem.*, 89(1):52–57, 1985.

[21] O. M. Becker and M. Karplus. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.*, 106(4):1495–1517, 1997.

[22] A. Ben-Naim and R. M. Mazo. Size dependence of the solvation free energies of large solutes. *J. Phys. Chem.*, 97:10829, 1993.

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

[24] R. S. Berry. Potential surfaces and dynamics: What clusters tell us. *Chem. Rev.*, 93(7):2379–2394, 1993.

[25] R. S. Berry, H. L. Davis, and T. L. Beck. Finding saddles on multidimensional potential surfaces. *Chem. Phys. Lett.*, 147(1):13–17, 1988.

[26] F. J. Blanco, G. Rivas, and L. Serrano. A short linear peptide that folds into a native stable beta-hairpin in aqueous solution. *Nat. Struct. Biol.*, 1(9):584–590, 1994.

[27] T.L. Blundell, B.L. Sibanda, M.J.E. Sternberg, and J.M. Thornton. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, 326:347, 1987.

[28] H. J. Böhm. The computer program ludi: A new method for the de novo design of enzyme inhibitors. *J. Comput. Aided Mol. Design*, 6:61–78, 1992.

[29] H. J. Böhm. Ludi: Rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput. Aided Mol. Design*, 6:593–606, 1992.

[30] A. M. J. J. Bonvin and W. F. van Gunsteren. β-hairpin stability and folding: Molecular dynamics studies of the first β-hairpin of tendamistat. *J. Mol. Biol.*, 296(1):255–268, 2000.

[31] 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(2):187–217, 1983.

[32] A.T. Brünger. X-PLOR, version 3.1 a system for x-ray crystallography and nmr. Yale University Press, New Haven, USA, 1992.

[33] S. K. Burley and G. A. Petsko. Aromatic–aromatic interaction: A mechanism of protein structure stabilization. *Science*, 229:23–28, 1985.

[34] B. D. Bursulaya and C. L. Brooks III. Folding free energy surface of a three-stranded β-sheet protein. *J. Am. Chem. Soc.*, 121(43):9947–9951, 1999.

[35] A. Calfisch, A. Miranker, and M. Karplus. Multiple copy simultaneous search and construction of ligands in binding sites: application of inhibitors of hiv-1 aspartic proteinase. *J. Med. Chem.*, 36:2142–2164, 1993.

[36] A. Calfisch, P. Niederer, and M. Anliker. Monte carlo docking of oligopeptides to proteins. *Proteins*, 13:223–230, 1992.

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

[38] E. Di Cera. *Thermodynamic Theory of Site-Specific Binding Processes in Biological Macromolecules.* Cambridge Univ. Press, 1995.

[39] C. J. Cerjan and W. H. Miller. On finding transition states. *J. Chem. Phys.*, 75(6):2800–2806, 1981.

[40] A. Chakrabartty and R. L. Baldwin. Stability of α-helices. *Adv. Protein Chem.*, 46:141–175, 1995.

[41] R. Chandrasekaran and G.N. Ramachandran. Studies on the conformation of amino acids, xi. analysis of the observed side group conformations in proteins. *Int. J. Protein Res.*, 2:223, 1970.

[42] J. Cherfils, S. Duquerroy, and J. Janin. Protein-protein recognition analyzed by docking simulation. *Proteins*, 11:271–280, 1991.

[43] S.Y. Chung and S. Subbiah. A structural explanation for the twilight zone of protein sequence homology. *Sructure*, 4:1123, 1996.

[44] B. W. Church, M. Orešič, and D. Shalloway. Tracking metastable states to free-energy global minima. In *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, volume 23, pages 41–64. American Mathematical Society, 1996.

[45] D. T. Clarke, A. J. Doig, B. J. Stapley, and G. R. Jones. The α-helix folds on the millisecond time scale. *Proc. Natl. Acad. Sci. USA*, 96(13):7232–7237, 1999.

[46] M. L. Connolly. Analytical molecular surface calculation. *J. Appl. Cryst.*, 16:548–558, 1983.

[47] M. L. Connolly. Solvent accessible surfaces of proteins and nucleic acids. *Science*, 221:709–713, 1983.

[48] M. L. Connolly. Shape complementarity at the hemoglobin $\alpha_1\beta_1$ subunit interface. *Biopolymers*, 25:1229–1247, 1986.

[49] P. Culot, G. Dive, V. H. Nguyen, and J. M. Ghuysen. A quasi-newton algorithm for first-order saddle-point location. *Theor. Chim. Acta*, 82(3-4):189–205, 1992.

[50] R. Czerminski and R. Elber. Reaction path study of conformational transitions in flexible systems: Applications to peptides. *J. Chem. Phys.*, 92(9):5580–5601, 1990.

[51] V. Daggett and M. Levitt. Realistic simulations of native-protein dynamics in solution and beyond. *Annu. Rev. Biophys. Biomol. Struct.*, 22:353–380, 1993.

[52] V. Daggett, A. J. Li, and A. R. Fersht. 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, 1998.

[53] P. Dauber-Osguthorpe, V. A. Roberts, D. J. Osguthorpe, J. Wolff, M. Genest, and A. T. Hagler. Structure and energetics of ligand binding to peptides: Escherichia coli dihydrofolate reductase–trimethoprim, a drug receptor system. *Proteins*, 4:31, 1988.

[54] E. de Alba, M. A. Jiménez, M. Rico, and J. L. Nieto. Conformational investigation of designed short linear peptides able to fold into β-hairpin structures in aqueous solution. *Fold. Des.*, 1(2):133–144, 1996.

[55] A. Dejaegere and M. Karplus. Analysis of coupling schemes in free energy simulations: A unified description of nonbonded contributions to solvation free energies. *J. Phys. Chem.*, 100:11148–11164, 1996.

[56] R. L. DesJarlais, G. L. Seibel, I. D. Kuntz, P. S. Furth, J. C. Alvarez, P. R. Ortiz de Montellano, D. L. Decamp, L. M. Babe, and C. S. Craik. Structure based design of nonpeptide inhibitors specific for the human immunodeficiency virus-1 protease. *Proc. Natl. Acad. Sci. USA*, 87:6644–6648, 1990.

[57] R. L DesJarlais, R. P. Sheridan, G. L. Seibel, J. S. Dixon, I. D. Kuntz, and R. Venkataraghavan. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J. Med. Chem.*, 31:722–729, 1988.

[58] R. Diamond. On the comparison of conformations using linear and quadratic transformations. *ACTA Cryst.*, page 1, 1976.

[59] K. A. Dill, A. T. Phillips, and J. B. Rosen. Cgu: an algorithm for molecular structure prediction. In *IMA Volumes in Mathematics and its Applications*, volume 94, pages 1–21. Springer-Verlag, 1997.

[60] K.A. Dill, A.T. Phillips, and J.B. Rosen. Protein structure and energy landscape dependence of sequence using a continuous energy function. *J. Computational Biology*, 4:227, 1997.

[61] A. R. Dinner, T. Lazaridis, and M. Karplus. Understanding beta-hairpin formation. *Proc. Natl. Acad. Sci. USA*, 96:9068–9073, 1999.

[62] 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.

[63] R.L. Dunbrack and M. Karplus. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *J. Mol. Biol.*, 230:543, 1993.

[64] H. J. Dyson and P. E. Wright. Defining solution conformations of small linear peptides. *A. Rev. Biophys. biophys. Chem.*, 20:519–538, 1991.

[65] W. A. Eaton, V. Munoz, P. A. Thompson, E. R. Henry, and J. Hofrichter. Kinetics and dynamics of loops, α-helices, β-hairpins and fast-folding proteins. *Acc. Chem. Res.*, 31(11):745–753, 1998.

[66] F. Eisenhaber and P. Argos. Improved strategy in analytic surface calculation for molecular systems: Handling of singularities and computational efficiency. *J. Comp. Chem.*, 14:1272–1280, 1993.

[67] F. Eisenhaber, P. Lijnzaad, P. Argos, C. Sander, and M. Scharf. The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *J. Comp. Chem.*, 16:273–284, 1995.

[68] R. Fletcher and M. J. D. Powell. A rapidly convergent descent method for minimization. *Computer Journal*, 6(2):163–168, 1963.

[69] P. J. Flory. Foundations of rotational isomeric state theory and general methods for generating configurational averages. *Macromolecules*, 7(3):381–392, 1974.

[70] C. A. Floudas. *Nonlinear and mixed-integer optimization.* Oxford University Press, New York, 1995.

[71] C. A. Floudas. *Deterministic Global Optimization: Theory, Methods and Applications.* Nonconvex Optimization and its Applications. Kluwer Academic Publishers, 2000.

[72] C.A. Floudas, J.L. Klepeis, and P.M. Pardalos. Global optimization approaches in protein folding and peptide docking. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 47:141–171, 1999.

[73] D. H. Fremont, W.A. Hendrickson, P. Marrack, and J. Kappler. Structures of an mhc class ii molecule with covalently bound single peptides. *Science*, 272:1001–1004, 1996.

[74] D. H. Fremont, M. Matsumura, E. A. Stura, P. A. Peterson, and I. A. Wilson. Crystal structures of two viral peptides in complex with murine mhc class i h-2 k^b. *Science*, 257:919–927, 1992.

[75] X. Fu, C. Bono, S. Woulfe, C. Swearingen, N. Summers, F. Sinigaglia, A. Sette, B. Schwartz, and R.W. Carr. Pocket 4 of the hla-dr molecule is a major determinant of t cell recognition of peptide. *Journal of Experimental Medicine*, 181:915–926, 1995.

[76] D. M. Gay. Sumsl, 1980. (FORTRAN source code).

[77] S. Gerschgorin. Über die abgrenzung der eigenwerte einer matrix. *Izv. Akad. Nauk SSSR, Ser. fiz.-mat.*, 6:749–754, 1931.

[78] P. Ghosh, M. Amaya, E. Mellins, and D.C. Wiley. The structure of an intermediate in class ii mhc maturation: Clip bound to hla-dr3. *Nature*, 378:457–462, 1995.

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

[80] M. K. Gilson and B. H. Honig. Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis. *Proteins: Structure, Function and Genetics*, 4:7–18, 1988.

[81] M. K. Gilson, K. A. Sharp, and B. H. Honig. Calculating the electrostatic potential of molecules in solution: Method and error assessment. *Journal of Computational Chemistry*, 9:327–335, 1988.

[82] 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(11):4751–4767, 1969.

[83] N. Go and H. A. Scheraga. On the use of classical statistical mechanics in the treatment of polymer chain conformations. *Macromolecules*, 9(4):535–542, 1976.

[84] D. Goldberg. *Genetic algorithms in search, optimization and machine learning.* Addison-Wesley, 1989.

[85] P. J. Goodford. A computational procedure for determining energetically favorable binding sites on biologically important molecules. *J. Med. Chem.*, 28:849–857, 1985.

[86] D. S. Goodsell and A. J. Olson. Automated docking of substrates to proteins by simulated annealing. *Proteins*, 8:195–202, 1990.

[87] W. H. Graham, E. S. Carter II, and R. P. Hicks. Conformational analysis of met–enkephalin in both aqueous solution and in the presence of sodium dodecyl sulfate micelles using multi-dimensional nmr and molecular modeling. *Biopolymers*, 32:1755–1764, 1992.

[88] J. Greenstadt. Variations on variable-metric methods. *Math. Comp.*, 24(109):1–22, 1970.

[89] P. D. J. Grootenhuis and P. A. Kollman. Crown ether-neutral molecule interactions studied by molecular mechanics and free energy perturbation calculations. near quantitative agreement between theory and experimental binding free energies. *J. Am. Chem. Soc.*, 111:4046–4051, 1989.

[90] K. Gulukota, S. Vajda, and C. DeLisi. Peptide docking using dynamic programming. *J. Comput. Chem.*, 17:418–428, 1996.

[91] P. Güntert, C. Mumenthaler, and K. Wüthrich. Torsion angle dynamics for nmr structure calculation with the new program dyana. *J. Mol. Biol.*, 273:283–298, 1997.

[92] J. Hammer, C. Bolin, D. Papadopoulos, J. Walsky, J. Higelin, W. Danho, F. Sinigaglia, and Z.A. Nagy. High-affinity binding of short peptides to major histocompatibility complex class ii molecules by anchor combinations. *Proc. Natl. Acad. Sci.*, 91:4456, 1994.

[93] U. H. Hansmann, M. Masuya, and Y. Okamoto. *Proc. Natl. Acad. Sci. USA*, 94:10652–10656, 1997.

[94] U. H. E. Hansmann and Y. Okamoto. Finite-size scaling of helix-coil transitions in poly-alanine studied by multicannonical simulations. *J. Chem. Phys.*, 110(2):1267–1276, 1999.

[95] T. N. Hart and R. J. Read. A multiple-start monte carlo docking method. *Proteins*, 13:206–222, 1992.

[96] T. N. Hart and R. J. Read. Multiple-start monte carlo docking of flexible ligands. In *The Protein Folding Problem and Tertiary Structure Prediction*, pages 71–108. Birkhäuser, 1994.

[97] R. D. Head, M. L. Smyte, T. I. Oprea, C. L. Waller, S. M. Green, and G. R. Marshall. Validate : a new method for receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.*, 118:3959–3969, 1996.

[98] D. Hinds and M. Levitt. Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.*, 243:668, 1994.

[99] L. Holm and C. Sander. Fast and simple monte-carlo algorithm for side-chain optimization in proteins: application to model building by homology. *Proteins: Sruct. Funct. Genet.*, 14:213, 1994.

[100] S. Honda, N. Kobayashi, and E. Munekata. Theormodynamics of a β-hairpin structure: Evidence for cooperative formation of folding nucleus. *J. Mol. Biol.*, 295(2):269–278, 2000.

[101] B. Honig, K. Sharp, and A. Yang. Macroscopic models of aqueous solutions: Biological and chemical applications. *J. Phys. Chem.*, 97:1101–1109, 1993.

[102] A. J. Hopfinger. Polymer-solvent interactions for homopolypeptides in aqueous solution. *Macromolecules*, 4:731–737, 1971.

[103] R. Horst and P. M. Pardalos, editors. *Handbook of Global Optimization.* Kluwer Academic Publishers, 1995.

[104] R. Horst and H. Tuy. *Global optimization: deterministic approaches.* Springer-Verlag, Berlin, 1993. 2nd. rev. edition.

[105] J.K. Hwang and W.F. Liao. Side-chain prediction by neural networks and simulated annealing optimization. *Protein Eng.*, 8:363, 1995.

[106] M.G. Ierapetritou, I.P. Androulakis, D.S. Monos, and C.A. Floudas. *Optimization in Computational Chemistry and Molecular Biology: Local and Global Approaches*, chapter Structure Prediction of Binding Sites of MHC Class II Molecules Based on the Crystal of HLA–DRB1 and Global Optimization, pages 157–189. Kluwer Academic Publishers, 2000.

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

[108] A. N. Jain and M. A. Murcko. Computational methods to predict binding free energy in ligand-receptor complexes. *J. Med. Chem.*, 38:4953–4967, 1995.

[109] J. Janin and S. J. Wodak. Reaction pathway for the quartenary structure change in hemoglobin. *Biopolymers*, 24:509–526, 1985.

[110] M. F. Jarrold. Introduction to statistical reaction rate theories. In H. Haberland, editor, *Clusters of Atoms and Molecules*, chapter 2.7, pages 163–186. Springer, Berlin, 1994.

[111] F. Jiang and S-H. Kim. "soft docking" : matching of molecular surface cubes. *J. Mol. Biol.*, 219:79–102, 1991.

[112] G. Jones, P. Willett, and R. C. Glen. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.*, 245:43–53, 1995.

[113] W. L. Jorgensen and J. Tirado-Rives. The opls potential functions for proteins. energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110:1657–1666, 1988.

[114] A. D. MacKerell Jr., D. Bashford, M. Bellott, R. L. Dunbrack Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102(18):3586–3616, 1998.

[115] R. S. Judson, Y. T. Tan, E. Mori, C. Melius, E. P. Jaeger, A. M Treasurywala, and A. Mathiowetz. Docking flexible molecules: a case study of three proteins. *J. Comput. Chem.*, 16:1405–1419, 1995.

[116] A. H. Juffer, F. Eisenhaber, S. J. Hubbard, D. Walther, and P. Argos. Comparison of atomic solvation parametric sets: Applicability and limitations in protein folding and binding. *Protein Sci.*, 4:2499, 1995.

[117] R. Friedman K. A. Sharp, A. Nicholls and B. Honig. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science*, 252:106, 1991.

[118] W. Kabsh. A solution for the best rotation to relate two sets of vectors. *ACTA Cryst.*, page 922, 1976.

[119] W. Kabsh. A discussion of the solution for the best rotation to relate two sets of vectors. *ACTA Cryst.*, page 827, 1978.

[120] 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), 1988.

[121] Y. K. Kang, G. Némethy, and H. A. Scheraga. Free energies of hydration of solute molecules 1. improvement of hydration shell model by exact computations of overlapping volumes. *J. Phys. Chem.*, 91:4105, 1987.

[122] Y. K. Kang, G. Némethy, and H. A. Scheraga. Free energies of hydration of solute molecules 2. application of the hydration shell model to nonionic organic molecules. *J. Phys. Chem.*, 91:4109, 1987.

[123] Y. K. Kang, G. Némethy, and H. A. Scheraga. Free energies of hydration of solute molecules 3. application of the hydration shell model to charged organic molecules. *J. Phys. Chem.*, 91:4118, 1987.

[124] S. K. Kearsley, D. J. Underwood, R. P. Sheridan, and M. D. Miller. Flexibases: a way to enhance the use of molecular docking methods. *J. Comput. Aided Mol. Design*, 8:565–582, 1994.

[125] P. S. Kim and R. L. Baldwin. Intermediates in the folding reactions of small proteins. *Annu. Rev. Biochem.*, 59:631–660, 1990.

[126] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[127] A. Kitao, F. Hirata, and N. Go. Effects of solvent on the conformation and the collective motions of a protein. 2. structure of hydration in melittin. *J. Phys. Chem.*, 97:10223–10230, 1993.

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

[129] J. L. Klepeis and C. A. Floudas. Comparative study of global minimum energy conformations of hydated peptides. *J. Computational Chemistry*, 20:636, 1999.

[130] J. L. Klepeis and C. A. Floudas. Ab-initio structure prediction in protein folding. 2000. In preparation.

[131] J. L. Klepeis and C. A. Floudas. Deterministic global optimization and torsion angle dynamics for molecular structure prediction. *Comp. Chem. Eng.*, 24:1761–1766, 2000.

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

[133] P. Koehl and M. Delarue. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.*, 239:249, 1994.

[134] P. Kollman. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.*, 93:2395–2417, 1993.

[135] T. Kortemme, M. Ramirez-Alvarado, and L. Serrano. Design of a 20-amino acid, three-stranded beta-sheet protein. *Science*, 281:253–256, 1998.

[136] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Landgridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161:269–288, 1982.

[137] R. E. Kunz and R. S. Berry. Statistical interpretation of topographies and dynamics of multidimensional potentials. *J. Chem. Phys.*, 103(5):1904–1912, 1995.

[138] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J. Mol. Bio.*, 157:105, 1982.

[139] M. C. Lawrence and P. C. Davis. Clix: A search algorithm for finding novel ligands capable of binding proteins of known three-dimensional structure. *Proteins*, 12:31–41, 1992.

[140] T. Lazaridis and M. Karplus. Effective energy function for proteins in solution. *Proteins: Struct. Funct. Genet.*, 35(2):133–152, 1999.

[141] A. R. Leach and I. D. Kuntz. Conformational analysis of flexible ligands in macromolecular receptor sites. *J. Comput. Chem.*, 13:733–748, 1992.

[142] B. Lee and F. M. Richards. The interpretation of protein structures : estimation of static accessibility. *J. Mol. Biol.*, 55:379–400, 1971.

[143] J. Y. 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. *Int. J. Quant. Chem.*, 75(3):255–265, 1999.

[144] J. Y. Lee, H. A. Scheraga, and S. Backovsky. New optimization method for conformational energy calculations on polypeptides: Conformational space annealing. *J. Comp. Chem.*, 18(9):1222–1232, 1997.

[145] J. Y. Lee, H. A. Scheraga, and S. Rackovsky. Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing. *Biopolymers*, 46:103–115, 1998.

[146] R. H. Lee and G. D. Rose. Molecular recognition. i. automatic identification of topographic surface features. *Biopolymers*, 24:1613–1627, 1985.

[147] P. Leopold, M. Montal, and J. Onuchic. Protein folding funnels : A kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci. USA*, 89:8721–8725, 1992.

[148] C. Levinthal. Are there pathways to protein folding ? *J. Chem. Phys.*, 65:44–45, 1968.

[149] M. Levitt. Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Biol.*, 170:723–764, 1983.

[150] Z. Li and H. A. Scheraga. Structure and free energy of complex thermodynamic systems. *J. Mol. Struct. (Theochem.)*, 179:333–352, 1988.

[151] J-H. Lii and N. L. Allinger. Molecular mechanics. the mm3 force field for hydrocarbons. 2. vibrational frequencies and thermodynamics. *J. Am. Chem. Soc.*, 111:8566–8575, 1989.

[152] J-H. Lii and N. L. Allinger. Molecular mechanics. the mm3 force field for hydrocarbons. 3. the van der waals' potentials and crystal data for aliphatic and aromatic hydroc arbons. *J. Am. Chem. Soc.*, 111:8576–8582, 1989.

[153] A. Liwo, J. Lee, D. Ripoll, J. Pillardy, and H.A. Scheraga. Surmounting the multiple minima problem in protein folding. *Proc. Natl. Acad. Sci. USA*, 96:5482, 1999.

[154] B. Ma and R. Nussinov. Molecular dynamics simulations of a β-hairpin fragment of protein g: Balance between side-chain and backbone forces. *J. Mol. Biol.*, 296(4):1091–1104, 2000.

[155] A. L. Mackay. The generalized inverse and inverse structure. *ACTA Cryst.*, page 212, 1977.

[156] V. Madison and K. D. Kopple. Solvent–dependent conformational distributions of some dipeptides. *J. Am. Chem. Soc.*, 102(15):4855–4863, 1980.

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

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

[159] C. D. Maranas and C. A. Floudas. Finding all solutions of non-linearly constrained systems of equations. *J. Global Opt.*, 7(2):143–182, 1995.

[160] S. Marqusee and R. L. Baldwin. Helix stabilization by $\text{glu}^{-} \cdots \text{lys}^{+}$ salt bridges in short peptides of de novo design. *Proc. Natl. Acad. Sci. USA*, 84:8898–8902, 1987.

[161] G. P. McCormick. Computability of global solutions to factorable nonconvex programs : Part i – convex underestimating problems. *Math. Programming*, 10:147–175, 1976.

[162] R. F. McGuire, F. A. Momany, and H. A. Scheraga. Energy parameters in polypeptides. v. an empirical hydrogen bond potential function based on molecular orbital calculations. *J. Phys. Chem.*, 76:375–393, 1972.

[163] A. D. McLachlan. A mathematical procedure for superimposing atomic coordinates of proteins. *ACTA Cryst.*, A28:656, 1972.

[164] A. D. McLachlan. Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.*, 128:49, 1979.

[165] H. Meirovitch and E. Meirovitch. Efficiency of monte carlo minimization procedures and their us e in analysis of nmr data obtained from flexible peptides. *J. Comput. Chem.*, 18:240–253, 1997.

[166] H. Meirovitch and M. Vásquez. Efficiency of simulated annealing and the monte carlo minimization method for generating a set of low energy structures of peptides. *J. Mol. Struct. (Theochem.)*, 398-399:517–522, 1997.

[167] E. C. Meng, B. K. Shoichet, and I. D. Kuntz. Automated docking with grid-based energy evaluation. *J. Comput. Chem.*, 13:505–524, 1992.

[168] M. D. Miller, S. K. Kearsley, D. J. Underwood, and R. P. Sheridan. Flog : A system to select "quasi-flexible" ligand complementary to a receptor of known three-dimensional structure. *J. Comput. Aided Mol. Design*, 8:153–174, 1994.

[169] A. Miranker and M. Karplus. Functionality maps of binding sites: a multicopy simultaneous search method. *Proteins*, 1:29–34, 1991.

[170] S. Miyamoto and P. A. Kollman. What determines the strength of noncovalent association of ligands to proteins in aqueous solutions ? *Proc. Natl. Acad. Sci. USA.*, 90:8402–8406, 1993.

[171] F. A. Momany, L. M. Carruthers, R. F. McGuire, and H. A. Scheraga. Intermolecular potential from crystal data. iii. *J. Phys. Chem.*, 78:1595–1620, 1974.

[172] F. A. Momany, L. M. Carruthers, and H. A. Scheraga. Intermolecular potential from crystal data. iv. *J. Phys. Chem.*, 78:1621–1630, 1974.

[173] F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga. Energy parameters in polypeptides. vii. *J. Phys. Chem.*, 79:2361–2381, 1975.

[174] A. Monge, R.A. Friesner, and B. Honig. An algorithm to generate low resolution protein tertiary structures from knowledge of secondary structure. *Proc. Natl. Acad. Sci. USA*, 91:5027, 1994.

[175] A. Monge, E.J.P. Lathrop, J.R. Gunn, P.S. Shenkin, and R. A. Friesner. Computer modeling of protein folding : Conformational and energetic analysis of reduced and detailed models. *J. Mol. Biology*, 247:995, 1995.

[176] J. B. Moon and J. Howe. Computer design of bioactive molecules: a method for receptor-based de novo ligand design. *Proteins*, 11:314–328, 1991.

[177] D. Morikis, N. Assa-Munt, A. Sahu, and J. D. Lambris. Solution structure of compstatin, a potent complement inhibitor. *Protein Sci.*, 7:619–627, 1998.

[178] V. Munoz, E. R. Henry, J. Hofrichter, and W. A. Eaton. A statistical mechanical model for β-hairpin kinetics. *Proc. Natl. Acad. Sci. USA*, 95(11):5872–5879, 1998.

[179] Bruce A. Murtagh and Michael A. Saunders. *MINOS 5.4 User's Guide.* Systems Optimization Laboratory, Department of Operations Research, Stanford University, 1993. Technical Report SOL 83-20R.

[180] 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: X. improved geometrical parameters and nonbonded interactions for use in the ecepp/3 algorithm, with application to proline-containing peptides. *J. Phys. Chem.*, 96(15):6472–6484, 1992.

[181] G. Némethy, M. S. Pottle, and H. A. Scheraga. Energy parameters in polypeptides. 9. *J. Phys. Chem.*, 87:1883–1887, 1983.

[182] A. Neumaier. *Interval Methods for Systems of Equations.* Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1990.

[183] T. Noguti and N. Go. A method of rapid calculation of a second derivative matrix of conformational energy for large molecules. *J. Phys. Soc. Japan*, 52(10):3685–3690, 1983.

[184] D. O'Neal, H. Taylor, and J. Simons. Potential surface walking and reaction paths for $be + h_2 \rightarrow beh_2 \rightarrow be + 2h$. *J. Phys. Chem.*, 88(8):1510–1513, 1984.

[185] T. Ooi, M. 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, 1987.

[186] 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*, Suppl. 3:149, 1999.

[187] A.R. Ortiz, A. Kolinski, and J. Skolnick. Native like topology assembly of small proteins using predicted restraints in monte carlo folding simulations. *Proc. Natl. Acad. Sci. USA*, 95:1020–1025, 1998a.

[188] C. M. Oshiro, I. D. Kuntz, and J. S. Dixon. Flexible ligand docking using a genetic algorithm. *J. Comput. Aided Mol. Design*, 9:113–130, 1995.

[189] V. S. Pande and D. S. Rokhsar. Molecular dynamics simulations of unfolding and refolding of a β-hairpin fragment of protein g. *Proc. Natl. Acad. Sci. USA*, 96(16):9062–9067, 1999.

[190] P. M. Pardalos, X. Liu, and G. L. Xue. Protein conformation of a lattice model using tabu search. *J. Global Optim.*, 11:55–68, 1997.

[191] G. Perrot, B. Cheng, K. D. Gibson, K. A. Palmer J. Vila, A. Nayeem, B. Maigret, and H. A. Scheraga. Mseed: A program for the rapid analytical determination of accessible surface areas and their derivatives. *J. Comp. Chem*, 13:1–11, 1992.

[192] J. W. Ponder. *TINKER, Software Tools for Molecular Design, Version 3.6.* Dept. of Biochemstry and Molecular Biophysics; Washington University School of Medicine; St. Louis, MO, 1998.

[193] M. Ramirez-Alvarado, F. J. Blanco, and L. Serrano. De novo design and structural analysis of a model beta-hairpin peptide system. *Nat. Struct. Biol.*, 3(7):604–612, 1996.

[194] S. T. Rao and M.G. Rossmann. *J. Mol. Biol.*, 76:241, 1973.

[195] H. Ratschek and J. Rokne. *Computer Methods for the Range of Functions.* Ellis Horwood Series in Mathematics and its Applications. Halsted Press, 1988.

[196] S.J. Remington and B.W. Matthews. *Proc. Nat. Acad. Sci. USA*, 75:2180, 1978.

[197] C. A. Reynolds, P. M. King, and W. G. Richards. Free energy calculations in molecular biophysics. *Mol. Phys.*, 76:251–275, 1992.

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

[199] D. Rogman, L. Scapozza, G. Folkers, and A. Daser. Molecular dynamics simulation oof mhc-peptide complexes as a tool for predicting potential t cell epitopes. *Biochemistry*, 33:11476–11485, 1994.

[200] R. Rosenfeld, Q. Zheng, S. Vajda, and C. DeLisi. Computing the structure of bound peptides. application to antigen recognition by class i major histocompatibility complex receptors. *J. Mol. Biol.*, 234:515–521, 1993.

[201] S. H. Rotstein and M. A. Murcko. Gensstar: a program for de novo drug design. *J. Comput. Aided Mol. Design*, 7:23–43, 1993.

[202] S. H. Rotstein and M. A. Murcko. Groupbuild: a fragment-based method for de novo drug design. *J. Med. Chem.*, 36:1700–1710, 1993.

[203] A. Sahu, B.K. Kay, and J.D. Lambris. Inhibition of human complement by a c3-binding peptide isolated from a phage-displayed random peptide library. *J. Immunol.*, 157:884–891, 1996.

[204] H. Schauber, F. Eisenhaber, and P. Argos. Rotamers: to be or not to be? an analysis of amino acid side-chain conformations in globular proteins. *J. Mol. Biol.*, 230:592, 1993.

[205] H.A. Scheraga. *PACK: Programs for Packing Polypeptide Chains*, 1996. online documentation.

[206] H.A. Scheraga, J. Lee, J. Pillardy, Y.J. Lee, A. Liwo, and D. Ripoll. Surmounting the multiple minima problem in protein folding. *J. Global Optimization*, 15:235, 1999.

[207] C. A. Schiffer, J. W. Caldwell, P. A. Kollman, and R. M. Stroud. Protein structure prediction with a combined solvation free energy-molecular mechanics force field. *Mol. Sim.*, 10:121, 1993.

[208] H. B. Schlegel. Geometry optimization on potential energy surfaces. In D. R. Yarkony, editor, *Modern Electronic Structure Theory*, chapter 8, pages 459–500. World Scientific Publishing, Singapore, 1995.

[209] K. A. Sharp and B. Honig. Electrostatic interactions in macromolecules: Theory and applications. *Annual Review of Biophysics and Biophysical Chemistry*, 19:301–332, 1990.

[210] F. B. Sheinerman and C. L. Brooks III. Calculations on folding of segment b1 of streptococcal protein g. *J. Mol. Biol.*, 278(2):439–456, 1998.

[211] J. Shen and F. A. Quiocho. Calculation of binding energy differences for receptor-ligand systems using the poisson-boltzmann methods. *J. Comput. Chem.*, 16:445–448, 1995.

[212] K. R. Shoemaker, P. S. Kim, E. J. York, J. M. Stewart, and R. L. Baldwin. Tests of the helix dipole model for stabilization of α-helices. *Nature*, 326:563–567, 1987.

[213] B. K. Shoichet and I. D. Kuntz. Protein docking and complementarity. *J. Mol. Biol.*, 221:327–346, 1991.

[214] D. Shortle, K.T. Simons, and D. Baker. Clustering of low energy conformations near the native structure of small proteins. *Proc. Natl. Acad. Sci. USA*, 95:11158, 1998.

[215] M. L. Silver, H. C. Guo, J. L. Strominger, and D. Wiley. Atomic structure of a human mhc molecule presenting an influenza virus peptide. *Nature*, 360:367–368, 1992.

[216] J. Simons, P. Jorgensen, H. Taylor, and J. Ozment. Walking on potential energy surfaces. *J. Phys. Chem.*, 87(15):2745–2753, 1983.

[217] K.T. Simons, I. Ruczinki, C. Kooperberg, B.A. Fox, C. Bystroff, and D. Baker. Improved recognition of native like structures using a combination of sequence dependent and sequence independent features of proteins. *Proteins*, 34:82, 1999a.

[218] J. Skolnick, A. Kolinski, and A.R. Ortiz. Monsster : a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.*, 265:217, 1997.

[219] R. Srinivasan and G. D. Rose. A physical basis for protein secondary structure. *PNAS*, 96:14258–14263, 1999.

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

[221] L. Stern, J. Brown, T. Jardetzky, J. Gorga, R. Urban, L. Strominger, and D. Wiley. Crystal structure of the human class ii mhc protein hla-dr1 complexed with an influenza virus peptide. *Nature*, 368:215–221, 1994.

[222] F. H. Stillinger and T. A. Weber. Inherent pair correlation in simple liquids. *J. Chem. Phys.*, 80(9):4434–4437, 1984.

[223] T. P. Straatsma and J. A. McCammon. Computational alchemy. *Annu. Rev. Phys. Chem.*, 43:407–435, 1992.

[224] L. Stryer. *Biochemistry.* W.H. Freeman and Company, fourth edition, 1995.

[225] S. Sun, P.D. Thomas, and K.A. Dill. A simple protein folding algorithm using a binary code and secondary structure constraints. *Protein Engineering*, 8:769, 1995.

[226] M.J. Sutcliffe, I. Haneef, D. Carney, and T.L. Blundell. Knowledge-based modeling of homologous proteins, part i: three dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.*, 1:377, 1987.

[227] Jean-Yves Trosset and Harold A. Scheraga. Prodock: Software package for protein modeling and docking. *Journal of Computational Chemistry*, 20:412–427, 1999.

[228] C. J. Tsai and K. D. Jordan. Use of an eigenmode method to locate the stationary points on the potential-energy surfaces of selected argon and water clusters. *J. Phys. Chem.*, 97(43):11227–11237, 1993.

[229] P. Tuffery, C. Etchebest, S. Hazout, and R. Lavery. A new approach to the rabid determination of protein side-chain conformations. *J. Biomol. Struct. Dynam.*, 8:1267, 1991.

[230] R. G. Urban and R. M. Chicz. *MHC Molecules: Expression, Assembly and Function.* R. G. Landes Company and Chapman & Hall, 1996.

[231] S. Vajda, M. Sippl, and J. Novotny. Empirical potentials and functions for protein folding and binding. *Curr. Opinion Struct. Biol.*, 7:222–228, 1997.

[232] W. F. van Gunsteren and H. J. C. Berendsen. *GROMOS.* Groningen Molecular Simulation, Groningen, The Netherlands, 1987.

[233] M. Vasquez. An evaluation of discrete and continuous search techniques for conformational analysis of side-chains in proteins. *Biopolymers*, 36:53, 1995.

[234] M. Vásquez, G. Némethy, and H. A. Scheraga. Conformational energy calculations on polypeptides and proteins. *Chemical Reviews*, 94:2183–2239, 1994.

[235] G. Verkhivker, K. Appelt, S. T. Freer, and J. E. Vilafranca. Empirical free energy calculations of ligand-protein crystallographic complexes. 1. knowledge based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Eng.*, 8:677–691, 1995.

[236] J. Vila, R.L. Williams, M. Vasquez, and H.A. Scheraga. Empirical solvation models can be used to differentiate native from non-native conformations of bovine pancreatic trypsin inhibitor. *Proteins*, pages 199–218, 1991.

[237] B. von Freyberg and W. Braun. Minimization of empirical energy functions in proteins including hydrophobic surface area effects. *J. Comp. Chem.*, 14:510–521, 1993.

[238] A. Šali, E. Shaknovich, and M. Karplus. Thermodynamics and kinetics of protein folding. In *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, volume 23, pages 199–213. American Mathematical Society, 1996.

[239] A. Wallquist, R. L. Jernigan, and D. G. Covell. A preference-based free energy parameterization of enzyme-inhibitor binding. applications to hiv-1 protease inhibitor design. *Protein Sci.*, 4:1881–1903, 1995.

[240] H. Wang. Grid-search molecular accessible surface algorithm for solving the protein docking problem. *J. Comput. Chem.*, 12:746–750, 1991.

[241] R. J. Wawak, K. D. Gibson, and H. A. Scheraga. Gradient discontinuities in calculations involving molecular surface area. *J. Math. Chem.*, 15:207–232, 1994.

[242] 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. A*, 102(17):2904–2918, 1998.

[243] S. Weiner, P. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106(3):765–784, 1984.

[244] S. Weiner, P. Kollman, D. Nguyen, and D. Case. An all atom force field for simulations of proteins and nucleic acids. *J. Comp. Chem.*, 7:230–252, 1986.

[245] L. Wesson and D. Eisenberg. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Science*, 1:227, 1992.

[246] K. M. Westerberg and C. A. Floudas. Dynamics of peptide folding: Transition states and reaction pathways of solvated and unsolvated tetra-alanine. *J. Global Opt.*, 15(3):261–297, 1999.

[247] K. M. Westerberg and C. A. Floudas. Locating all transition states and studying the reaction pathways of potential energy surfaces. *J. Chem. Phys.*, 110(18):9259–9295, 1999.

[248] R. L. Williams, J. Vila, G. Perrot, and H. A. Scheraga. Empirical solvation models in the context of conformational energy searches: Application to bovine pancreatic trypsin inhibitor. *Proteins*, 14:110–119, 1992.

[249] S. J. Wodak, M. De Crombrugghe, and J. Janin. Computer studies of interactions between macromolecules. *Prog. Biophys. Mol. Biol.*, 49:29–63, 1987.

[250] S. J. Wodak and J. Janin. Computer analysis of protein-protein interactions. *J. Mol. Biol.*, 124:323–342, 1978.

[251] R. Wolfenden, L. Andersson, P. M. Cullis, and C. C. B. Southgate. Affinities of amino acid side chains for solvent water. *Biochem.*, 20:849, 1981.

[252] K. Yue and K.A. Dill. Folding proteins with a simple energy function and extensive conformational searching. *Protein Science*, 5:254, 1996.

[253] S. S. Zimmerman, M. S. Pottle, G. Némethy, and H. A. Scheraga. Conformational analysis of the 20 naturally occurring amino acid residues using ecepp. *Macromolecules*, 10:1–9, 1977.