Chapter 1

DETERMINISTIC GLOBAL OPTIMIZATION FOR PROTEIN STRUCTURE PREDICTION

John L. Klepeis
Department of Chemical Engineering
Princeton University
Princeton, NJ 08544
john@titan.princeton.edu

Christodoulos A. Floudas
Department of Chemical Engineering
Princeton University
Princeton, NJ 08544
floudas@titan.princeton.edu

Abstract Deterministic global optimization plays an essential role in the solution of many difficult problems with applications ranging from economics and operations research to computational chemistry and molecular biology. In this chapter we explore the application of deterministic global optimization approaches to problems related to protein structure prediction. Due to the complex nature of protein interactions, energy landscapes which model these systems display huge numbers of local minima often separated by high energy barriers. Since the number of local minima is vast, the corresponding formulation has earned the simple yet suggestive title of "multiple-minima" problem. Based on the complexity of the energy hypersurface, there is an obvious need for the development of effective global optimization techniques. In this work, we have focused on the development of such global optimization methods through the foundations of the αBB deterministic global optimization approach.

Keywords: Deterministic global optimization, Protein folding, Structure prediction, Energy modeling

1. INTRODUCTION

Proteins are undoubtedly the most complex and vital molecules in nature. This complexity arises from an intricate balance of intra- and inter-molecular interactions which define the native three-dimensional structure of the system, and subsequently its biological functionality. Recent advances in genetic engineering have heightened the interest in research related to understanding the dynamics and predicting the equilibrium native protein folding and docking conformations. The ability to predict these structures is of great theoretical interest, especially in the fields of biophysics and biochemistry. Moreover, the applications of such knowledge also promise to be exciting. For example, the ability to predict these structures would greatly increase our understanding of hereditary and infectious diseases and aid in the interpretation of genome data. Such knowledge would also likely revolutionize the process of de novo drug design.

Anfinsen's thermodynamic hypothesis (Anfinsen et al., 1961) suggests that this native structure is in a state of thermodynamic equilibrium corresponding to the system with the lowest free energy. Experimental studies have since shown that, under native physiological conditions and after denaturation, globular proteins spontaneously refold to their unique, native structure (Kim and Baldwin, 1990). Understanding the transition of a protein from a disordered state to its native state defines the protein folding problem.

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 "multiple-minima" problem. The basis for these difficulties is best summarized by Levinthal's paradox (Levinthal, 1968). 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 (Becker and Karplus, 1997; Czerminski and Elber, 1990; Church et al., 1996; Leopold et al., 1992; Šali et al., 1996). 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.

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. This is followed by a comprehensive study of ab-initio modeling for structure prediction of single chain polypeptides. 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.

2. THEORY

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

$$\begin{array}{rr} \min\limits_{\mathbf{x}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{g}(\mathbf{x}) \leq\ 0 \\ & \mathbf{h}(\mathbf{x}) =\ 0 \\ & \mathbf{x} \in [\mathbf{x}^L, \mathbf{x}^U] \end{array} \tag{1.1}$$

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 (1.1), namely, stochastic and deterministic approaches. Stochastic methods, such as those based on genetic algorithms (Goldberg, 1989) and simulated annealing (Kirkpatrick et al., 1983), 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 (Floudas, 2000; Horst and Pardalos, 1995; Horst and Tuy, 1993).

The αBB algorithm has been developed to address general twice continuously differentiable models of type 1.1 (Adjiman and Floudas, 1996; Adjiman et al., 1996; Adjiman et al., 1998b; Adjiman et al., 1998a; Androulakis et al., 1995). The algorithm is built on a branch-and-bound framework and can handle generic nonconvex optimization problems represented by formulation (1.1). 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 (Adjiman et al., 1998b; Adjiman et al., 1998a; Maranas and Floudas, 1994a; Maranas and Floudas, 1994b).

The αBB global optimization approach is based on the convex relaxation of the original nonconvex formulation (1.1). 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 (Androulakis et al., 1995).

When applying the αBB approach to the protein folding problem, formulation (1.1) 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.

2.1 UNDERESTIMATING TERMS OF SPECIAL STRUCTURE

In the case of a bilinear term xy, (Al-Khayyal and Falk, 1983) 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:

$$w_B = \max\{x^L y + y^L x - x^L y^L;\ x^U y + y^U x - x^U y^U\}. \tag{1.2}$$

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

$$\begin{array}{lcl} w_B & \geq & x^L y + y^L x - x^L y^L, \\ w_B & \geq & x^U y + y^U x - x^U y^U. \end{array} \tag{1.3}$$

Moreover, an upper bound can be imposed on w to construct a better approximation of the original problem (McCormick, 1976). This is achieved through the addition of two linear constraints:

$$\begin{array}{lcl} w_B & \leq & x^U y + y^L x - x^U y^L, \\ w_B & \leq & x^L y + y^U x - x^L y^U. \end{array} \tag{1.4}$$

A trilinear term of the form xyz can be underestimated in a similar fashion (Maranas and Floudas, 1995). A new variable w_T is introduced and bounded by the following eight inequality constraints:

$$\begin{array}{lcl} w_T & \geq & xy^L z^L + x^L y z^L + x^L y^L z - 2x^L y^L z^L, \\ w_T & \geq & xy^U z^U + x^U y z^L + x^U y^L z - x^U y^L z^L - x^U y^U z^U, \\ w_T & \geq & xy^L z^L + x^L y z^U + x^L y^U z - x^L y^U z^U - x^L y^L z^L, \\ w_T & \geq & xy^U z^L + x^U y z^U + x^L y^U z - x^L y^U z^L - x^U y^U z^U, \\ w_T & \geq & xy^L z^U + x^L y z^L + x^U y^L z - x^U y^L z^U - x^L y^L z^L, \\ w_T & \geq & xy^L z^U + x^L y z^U + x^U y^U z - x^L y^L z^U - x^U y^U z^U, \\ w_T & \geq & xy^U z^L + x^U y z^L + x^L y^L z - x^U y^U z^L - x^L y^L z^L, \\ w_T & \geq & xy^U z^U + x^U y z^U + x^U y^U z - 2x^U y^U z^U. \end{array} \tag{1.5}$$

Fractional terms of the form x/y are underestimated by introducing a new variable w_F and two new constraints (Maranas and Floudas, 1995) which depend on the sign of the bounds on x.

$$
\begin{aligned}
w_F &\geq \begin{cases} x^L/y + x/y^U - x^L/y^U & \text{if } x^L \geq 0 \\ x/y^U - x^L y/y^L y^U + x^L/y^L & \text{if } x^L < 0 \end{cases} \\
w_F &\geq \begin{cases} x^U/y + x/y^L - x^U/y^L & \text{if } x^U \geq 0 \\ x/y^L - x^U y/y^L y^U + x^U/y^U & \text{if } x^U < 0 \end{cases}
\end{aligned}
\tag{1.6}
$$

For fractional trilinear terms, eight new constraints are required (Maranas and Floudas, 1995). 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

$$
\begin{aligned}
&w_{FT} \geq xy^L/z^U + x^L y/z^U + x^L y^L/z - 2x^L y^L/z^U, \\
&w_{FT} \geq xy^L/z^U + x^L y/z^L + x^L y^U/z - x^L y^U/z^L - x^L y^L/z^U, \\
&w_{FT} \geq xy^U/z^L + x^U y/z^U + x^U y^L/z - x^U y^L/z^U - x^U y^U/z^L, \\
&w_{FT} \geq xy^U/z^U + x^U y/z^L + x^L y^U/z - x^L y^U/z^U - x^U y^U/z^L, \\
&w_{FT} \geq xy^L/z^U + x^L y/z^L + x^U y^L/z - x^U y^L/z^L - x^L y^L/z^U, \\
&w_{FT} \geq xy^U/z^U + x^U y/z^L + x^L y/z - x^L y^U/z^U - x^U y^U/z^L, \\
&w_{FT} \geq xy^L/z^U + x^L y/z^L + x^U y^L/z - x^U y^L/z^L - x^L y^L/z^U, \\
&w_{FT} \geq xy^U/z^L + x^U y/z^L + x^U y^U/z - 2x^U y^U/z^L.
\end{aligned}
\tag{1.7}
$$

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:

$$
ut(x^L) + \frac{ut(x^U) - ut(x^L)}{x^U - x^L}(x - x^L). \tag{1.8}
$$

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

2.2 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

$$\hat{f}(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^{n} \alpha_i (x_i^L - x_i)(x_i^U - x_i) \tag{1.9}$$

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 :

$$\max_{\mathbf{x}^L \leq \mathbf{x} \leq \mathbf{x}^U} \left[f(\mathbf{x}) - \hat{f}(\mathbf{x})\right] \;=\; \frac{1}{4} \sum_{i}^{n} \alpha_i \left(x_i^U - x_i^L\right)^2 . \tag{1.10}$$

(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 $f(\mathbf{x})$.

$$\alpha \;\geq\; \max\left\{0, -\frac{1}{2} \min_{i, \mathbf{x}^L \leq \mathbf{x} \leq \mathbf{x}^U} \lambda_i(\mathbf{x})\right\} \tag{1.11}$$

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 (Adjiman et al., 1996; Adjiman and Floudas, 1996). 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 (Ratschek and Rokne, 1988; Neumaier, 1990). The interval Hessian matrix family $[H_f]$ is then used to formulate a theorem in which the α calculation problem is relaxed (Adjiman et al., 1996). In other words, a valid lower bound on the minimum eigenvalue can be used to calculate rigorous α values :

$$\alpha \geq \left\{0, -\frac{1}{2}\lambda_{min}\left([H_f]\right)\right\} \tag{1.12}$$

where $\lambda_{min}\left([H_f]\right)$ 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 (Gerschgorin, 1931) 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

$$\lambda_{min} = \min_i \left(a_{ii} - \sum_{j \neq i} |a_{ij}|\right). \tag{1.13}$$

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 (1.9). 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.

2.3 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 (1.9), to yield the relaxed convex inequality $\hat{g}(\mathbf{x}) \leq 0$.

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,

$$\begin{array}{rrl} h^+(\mathbf{x}) = & h(\mathbf{x}) & \leq 0 \\ h^-(\mathbf{x}) = & -h(\mathbf{x}) & \leq 0. \end{array} \tag{1.14}$$

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

2.4 CONVEX LOWER BOUNDING PROBLEM FORMULATION

Summarizing the concepts introduced so far, a convex relaxation for any nonconvex problem of type (1.1) belonging to the broad class of twice continuously differentiable continuous NLPs can be constructed as

$$
\begin{array}{ll}
\min\limits_{\mathbf{x}} & \hat{f}(\mathbf{x}) \\
\text{subject to} & \hat{\mathbf{g}}(\mathbf{x}) \le 0 \\
& \hat{\mathbf{h}}^{+}(\mathbf{x}) \le 0 \\
& \hat{\mathbf{h}}^{-}(\mathbf{x}) \le 0 \\
& \mathbf{x} \in [\mathbf{x}^L, \mathbf{x}^U]
\end{array}
\tag{1.15}
$$

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 non-convex 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.

2.5 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

$$
x_i^{L,NEW} / x_i^{U,NEW} = \left\{
\begin{array}{ll}
\min\limits_{\mathbf{x}} / \max\limits_{\mathbf{x}} & x_i \\
\text{subject to} & \hat{\mathbf{g}}(\mathbf{x}) \le 0 \\
& \mathbf{x}^L \le \mathbf{x} \le \mathbf{x}^U
\end{array}
\right.
\tag{1.16}
$$

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.

2.6 THE αBB ALGORITHM

The global optimization method αBB deterministically locates the global minimum solution of (1.1) based on the refinement of converging lower and upper bounds. The lower bounds are obtained by the solution of (1.15), which is formulated as a convex programming problem. Upper bounds are based on the solution of (1.1) 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 (1.1) and its convex relaxation in (1.15) 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 (1.1) and its convex relaxation (1.15) to become less than ϵ. Therefore, any feasible point $\mathbf{x}$ of problem (1.15), including the global minimum solution, becomes at least ϵ–feasible for problem (1.1) 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 (1.1) 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 (1.1) is simply the minimum over all the minima of problem (1.15) 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 (1.15) 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 (1.1) locally and selecting it to be the minimum over all the previously recorded upper bounds. Obviously, if the single minimum of (1.15) 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).

Figure 1.1 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 (Maranas and Floudas, 1994a). In addition to computational chemistry related problems, the αBB approach has been applied to a variety of constrained optimization problems (Adjiman et al., 1998b; Adjiman et al., 1998a; Adjiman et al., 1996; Androulakis et al., 1995).

3. 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 through global optimization. In our work, we have focused both on the development of global optimization methods and 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.

3.1 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

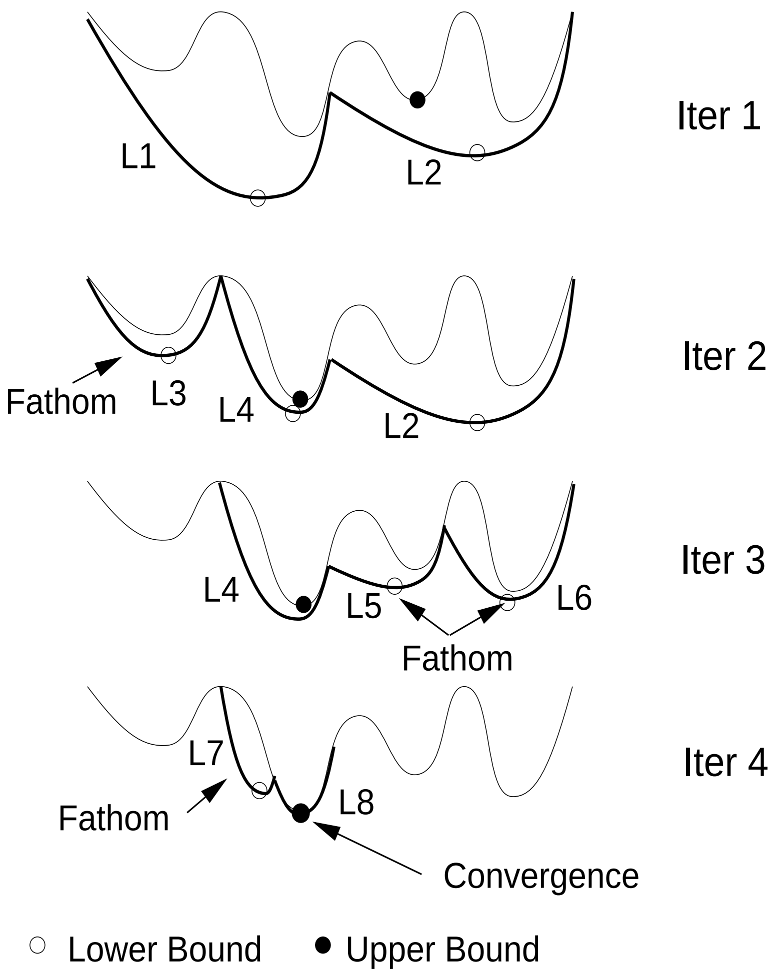

Figure 1.1 One-dimensional illustrative example of the αBB approach. In iteration 1 the overall domain is bisected, the two convex lower bounding functions are created and their unique minima ($L1$ and $L2$) are identified. An upper bound is also identified. Since $L1$ is less than $L2$, the region containing $L1$ 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.

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.

For this work the ECEPP/3 (Empirical Conformational Energy Program for Peptides) (Némethy et al., 1992) 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. 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 1.2, while the development of the appropriate parameters is discussed and reported elsewhere (Némethy et al., 1992).

3.2 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.

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

Figure 1.2 Potential energy terms in ECEPP/3 force field. r_{ij} refers to the interatomic distance of the atomic pair (ij). Q_i and Q_j are dipole parameters for the respective atoms, in which the dielectric constant of 2 has been incorporated. F_{ij} is set equal to 0.5 for 1–4 interactions and 1.0 for 1–5 and higher interactions. A_{ij}, C_{ij}, A'_{ij} and B_{ij} are nonbonded and hydrogen bonded parameters specific to the atomic pair. $E_{o,k}$ 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 $\AA$

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 (Dejaegere and Karplus, 1996; Kollman, 1993; Straatsma and McCammon, 1992). The integral equation method has also been used to analyze the solvent structure of a protein system (Kitao et al., 1993). 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 (Honig et al., 1993). Other continuum models estimate free energies of solvation as a function of surface areas and volumes.

In our work, solvation contributions are included implicitly using empirical correlations with both surface area (Perrot et al., 1992) and volume (Augspurger and Scheraga, 1996). 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:

$$E_{HYD} = \sum_{i=1}^{N} (A_i)(\sigma_i) \tag{1.17}$$

In Equation (1.17), 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. Once the solvent–accessible surface areas have been calculated, these values must be multiplied by the appropriate (σ_i) parameters as shown in Equation (1.17). A variety of atomic solvation parameter (ASP) sets have been developed to model the transfer of atoms from a gaseous to a hydrated environment. In our work, five ASP sets, namely the WE1 (Wesson and Eisenberg, 1992), WE2 (Wesson and

Eisenberg, 1992), OONS (Ooi et al., 1987), SCKS (Schiffer et al., 1993), and JRF (Williams et al., 1992) parameters, have been studied.

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:

$$E_{HYD} = \sum_{i=1}^{N} (VHS_i)(\delta_i) \tag{1.18}$$

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. Free energy density parameters for solvent accessible volumes have been developed for nonionic and charged organic solute molecules (Kang et al., 1987a; Kang et al., 1987b; Kang et al., 1988). 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 (Augspurger and Scheraga, 1996), are used.

3.3 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 1.1. 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:

$$\min \quad E(\phi_i, \psi_i, \omega_i, \chi_i^k, \theta_j^N, \theta_j^C) \tag{1.19}$$

$$\begin{aligned} \text{subject to} \quad -\pi &\leq \phi_i \leq \pi, \quad i = 1, \ldots, N_{RES} \\ -\pi &\leq \psi_i \leq \pi, \quad i = 1, \ldots, N_{RES} \\ -\pi &\leq \omega_i \leq \pi, \quad i = 1, \ldots, N_{RES} \end{aligned}$$

$$
\begin{aligned}
-\pi &\leq \chi_i^k \leq \pi, \ i = 1, \ldots, N_{RES}, \ k = 1, \ldots, K^i \\
-\pi &\leq \theta_j^N \leq \pi, \ j = 1, \ldots, J_N \\
-\pi &\leq \theta_j^C \leq \pi, \ j = 1, \ldots, J_C
\end{aligned}
$$

In general, E represents the total potential energy function and the free energy of solvation. However, in the case of one ASP set, in particular the JRF ASP (Williams et al., 1992), 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:

$$E_{JRF}^{Total} = E_{Min}^{Unsol} + E_{JRF}^{Sol} \tag{1.20}$$

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 (Adjiman et al., 1997; Adjiman et al., 1998b; Adjiman et al., 1998a; Adjiman et al., 1996; Androulakis et al., 1995) 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 determination of the global minimum energy conformation using αBB requires the interfacing of a number of programs (αBB (Adjiman et al., 1997; Adjiman et al., 1998b; Adjiman et al., 1998a; Adjiman et al., 1996; Androulakis et al., 1995), PACK (Scheraga, 1996), NPSOL (Gill et al., 1986) 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 (Gill et al., 1986), 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 (Perrot et al., 1992); whereas volumes of hydration shells are determined using the RRIGS module (Augspurger and Scheraga, 1996). 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 1.3.

4. 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 (Stillinger and Weber, 1984). However, a drawback of these approaches is their inherent stochastic nature. In its original form, the αBB *deterministic* global optimization algorithm (Adjiman et al., 1996; Adjiman et al., 1997; Adjiman et al., 1998b; Adjiman et al., 1998a; Androulakis et al., 1995) has been shown to be an efficient method for finding the global minimum energy conformation for both unsolvated and solvated peptide systems (Androulakis et al., 1997; Klepeis et al., 1998; Klepeis and Floudas, 1999). 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.

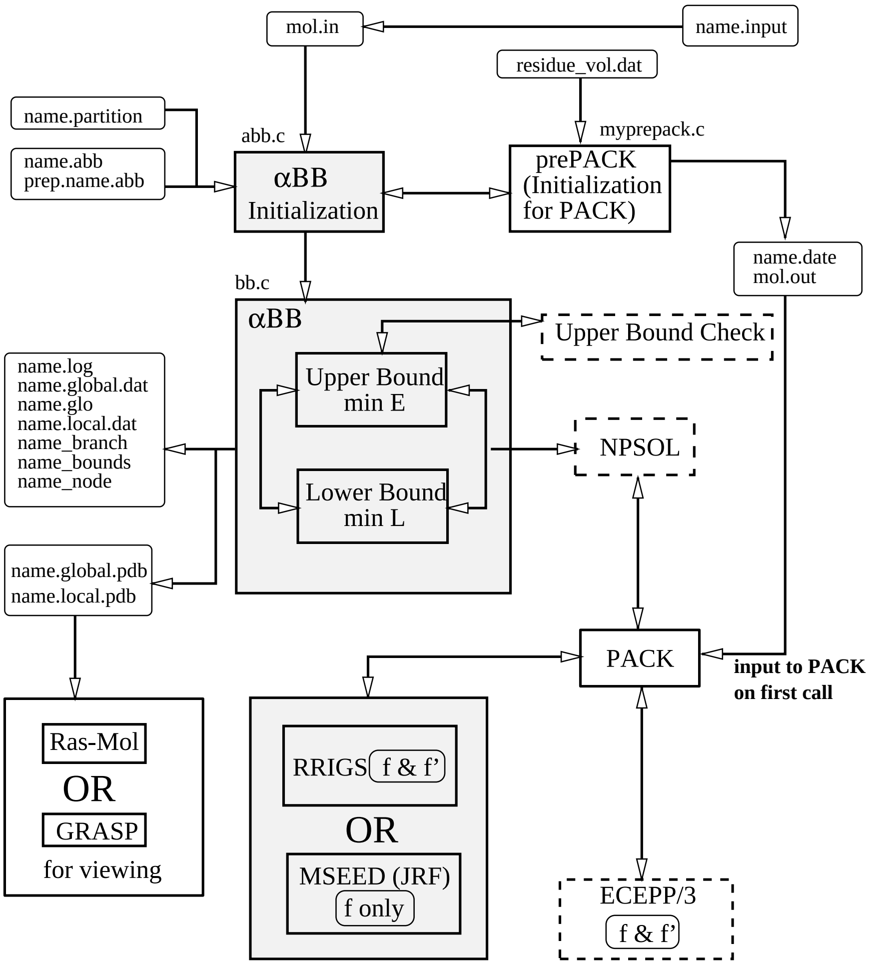

Figure 1.3 Interface for αBB within GLO-FOLD. The arrows indicate the direction of information flow. The names of the input, output, and intermediate files are indicated, in addition to selected source code files. References to "f & f'" and "f only" describe whether gradient evaluations or only function evaluations are used in the respective modules.

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.

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 :

$$p_\gamma^{exact} = \frac{\sum_{i(\gamma)\in I(\gamma)} exp(-\beta E_{i(\gamma)})}{\sum_{i\in I} exp(-\beta E_i)} \tag{1.21}$$

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

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 (Go and Scheraga, 1969; Go and Scheraga, 1976; Flory, 1974). In the work of (Go and Scheraga, 1969) (Go and Scheraga, 1969) a classical rigid model was used to characterize a partition function based on the fixed bond length and bond angle assumptions. In contrast, (Flory, 1974) (Flory, 1974) derived a different partition function using a classical flexible model. Later analysis by (Go and Scheraga, 1976) (Go and Scheraga, 1976) 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).

An approximate probability associated with a given basin (γ) can be calculated using the following equation :

$$p_{\gamma}^{approx} = \frac{\left[\left[Det\left(H_{\gamma}\right)\right]\right]^{-1/2} \exp\left(-\beta E_{\gamma}^{*}\right)}{\sum_{i=1}^{N}\left[\left[Det\left(H_{i}\right)\right]\right]^{-1/2} \exp\left(-\beta E_{i}^{*}\right)} \tag{1.22}$$

To develop a meaningful comparison of relative free energies, the total partition function (i.e., the denominator of Equation (1.22)) 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 (1.22)) :

$$P_{j}^{approx} = \sum_{\gamma=1}^{n_j} p_{\gamma}^{approx} \tag{1.23}$$

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 :

$$< E >_{T} = \sum_{j} P_{j}^{approx} < E >_{j} \tag{1.24}$$

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

4.1 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 (1.19). The new formulation is :

$$\min \quad E(\phi_i, \psi_i, \omega_i, \chi_i^k, \phi_j^N, \phi_j^C) \tag{1.25}$$

$$\begin{array}{rrcll}
\text{subject to} & & & (E^* - E) + \epsilon^* \; < \; 0 & \\
-\pi & \leq & \phi_i & \leq \; \pi, & i = 1, \ldots, N_{RES} \\
-\pi & \leq & \psi_i & \leq \; \pi, & i = 1, \ldots, N_{RES} \\
-\pi & \leq & \omega_i & \leq \; \pi, & i = 1, \ldots, N_{RES} \\
-\pi & \leq & \chi_i^k & \leq \; \pi, & i = 1, \ldots, N_{RES}, \; k = 1, \ldots, K^i \\
-\pi & \leq & \phi_j^N & \leq \; \pi, & j = 1, \ldots, J^N \\
-\pi & \leq & \phi_j^C & \leq \; \pi, & j = 1, \ldots, J^C
\end{array}$$

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 (1.19) is actually a special case of this problem in which $E^* = -\infty$ and $\epsilon^* = 0$. That is, in (1.19) 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 (Maranas and Floudas, 1995). 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 :

$$\frac{\partial E(\theta)}{\partial \theta_i} = 0 , \qquad i = 1, \ldots, N_\theta \tag{1.26}$$

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 (1.26) for which the Hessian of E is positive definite.

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.

4.2 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} (Li and Scheraga, 1988)), 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.

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 (Vásquez et al., 1994), and has been shown to be effective in locating low energy conformers of peptides (Meirovitch and Meirovitch, 1997; Meirovitch and Vásquez, 1997).

The problem formulation is identical to the one given in (1.19). 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:

$$G = U_{Min} + \frac{1}{2\beta} \ln \left[Det\left(H_{Min} \right) \right] \tag{1.27}$$

U_{Min} represents the local minimum energy of E or L, and $Det\,(H_{Min})$ 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 (1.27) must be specified as an additional input parameter.

4.3 FREE ENERGY COMPUTATIONAL STUDIES

The EDA was first applied to the isolated form of met-enkephalin. 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. The results suggest that the inclusion of entropic contributions greatly affects the relative stability of individual low energy structures.

The EDA was also 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 1.1, the PEGM structure persists 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 (Klepeis et al., 1998). The sequence of FEGM structures is illustrated in Figure 1.4.

5. 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, such as the computational approaches outlined above. Experimentally, there 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.

Table 1.1 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

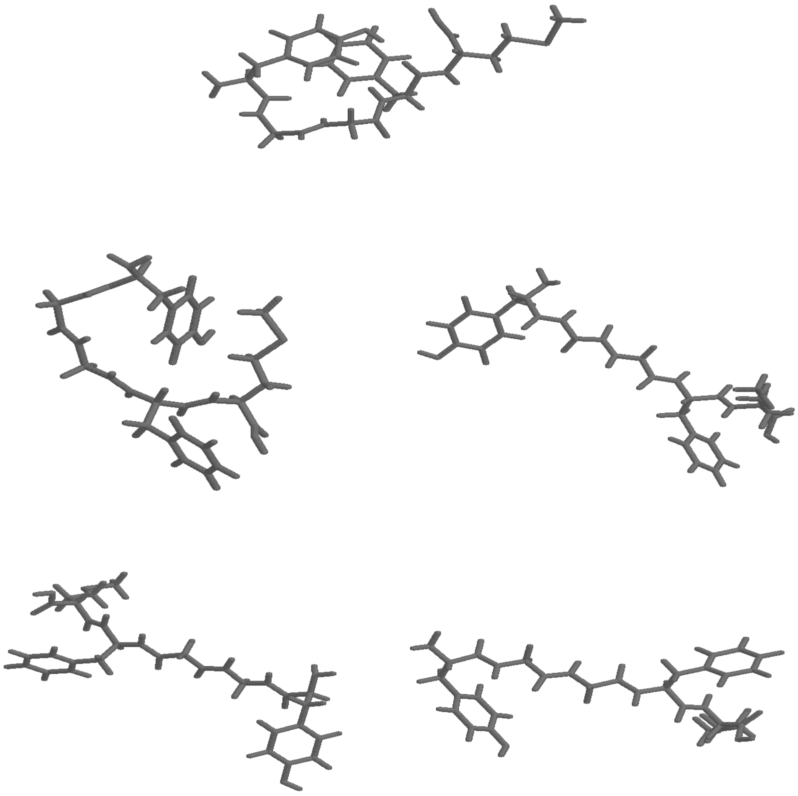

Figure 1.4 FEGM structures for solvated met-enkephalin. The top figure is the PEGM and the FEGM for 100 K. The structures at other temperatures (200,300,400,500) are shown left to right, top to bottom.

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 (Adjiman et al., 1996; Adjiman et al., 1997; Adjiman et al., 1998b; Adjiman et al., 1998a; Androulakis et al., 1995). αBB based global optimization techniques have also been applied to NMR type structure prediction problems (Klepeis et al., 1999; Standley et al., 1999).

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 (Güntert et al., 1997; Rice and Brünger, 1994). 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.

5.1 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:

$$E = E_{\text{forcefield}} + W_{\text{nmr}} E_{\text{nmr}}. \tag{1.28}$$

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 (1.28) can be rewritten as :

$$E_{\text{nmr}} = E_{\text{distance}} + E_{\text{dihedral}}. \tag{1.29}$$

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, $E_{\text{distance}} = E^{\text{upper}}_{\text{distance}} + E^{\text{lower}}_{\text{distance}}$. When considering upper distance restraints this becomes:

$$E^{\text{upper}}_{\text{distance}} = \sum_j \begin{cases} A_j(d_j - d_j^{\text{upper}})^2 & \text{if } d_j > d_j^{\text{upper}}, \\ 0 & \text{otherwise.} \end{cases} \tag{1.30}$$

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}:

$$E^{\text{lower}}_{\text{distance}} = \sum_j \begin{cases} A_j(d_j - d_j^{\text{lower}})^2 & \text{if } d_j < d_j^{\text{lower}}, \\ 0 & \text{otherwise.} \end{cases} \tag{1.31}$$

For dihedral angle restraints the functional form is similar to that of Equations (1.30) and (1.31). 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., $\Delta HW_{\omega_j} = \pi - (\omega_j^{\text{upper}} - \omega_j^{\text{lower}})/2$).

The force field energy term, $E_{\text{forcefield}}$ of Equation (1.28), 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 :

$$E_{\text{S}} = E_{\text{distance}} + E_{\text{dihedral}}. \tag{1.32}$$

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 (Némethy et al., 1992). When considering an unconstrained minimization, this approach provides the following objective function :

$$E_{\text{D}} = E_{\text{distance}} + E_{\text{dihedral}} + E_{\text{ECEPP/3}}. \tag{1.33}$$

In contrast to Equation (1.32), 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.

5.2 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 :

$$\min_{\phi} \qquad E_{\mathrm{S}}(\phi) = E_{\mathrm{distance}} + E_{\mathrm{dihedral}}. \tag{1.34}$$

A standard procedure for addressing this global optimization problem consists of a combination of simulated annealing and molecular or torsional angle dynamics (Brünger, 1992). 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 (1.33), the unconstrained formulation is similar to formulation (1.34). 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 $E_{\mathrm{forcefield}}$:

$$\min_{\phi} \qquad E_{\mathrm{ECEPP/3}}, \tag{1.35}$$

$$\begin{array}{llll} \text{subject to} & E_l^{\mathrm{distance}}(\phi) & \leq \ E_l^{\mathrm{ref}} & l = 1, \ldots, N_{\mathrm{CON}}, \\ & \phi_i^L & \leq \phi_i \leq \ \phi_i^U, & i = 1, \ldots, N_\phi. \end{array}$$

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.

5.3 TORSION ANGLE DYNAMICS

Standard unconstrained molecular dynamics simulations have been used extensively to model the folding and unfolding of protein systems (Duan and Kollman, 1998; Daggett et al., 1998; Caves et al., 1998).

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

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

5.4 COMPUTATIONAL STUDY

The global optimization algorithm was tested on Compstatin, a synthetic 13-residue (ICVVQDWGHHRCT) cyclic peptide (disulfide bridge between the Cys^2 and Cys^{12} residues) that binds to C3 (third component of complement) and inhibits complement activation (Sahu et al., 1996). Two-dimensional NMR techniques (Morikis et al., 1998) 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 (Morikis et al., 1998), traditional distance geometry–simulated annealing protocol was utilized to minimize the associated target function in the Cartesian coordinate space using the program X-PLOR (Brünger, 1992). 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 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. 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 1.5.

5.5 GLOBAL OPTIMIZATION AND TORSION ANGLE DYNAMICS

A modified constrained global optimization was also applied to the Compstatin structure prediction problem using the same constraint function and parameters (Klepeis and Floudas, 2000). 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 (Güntert et al., 1997).

To gauge the performance of the combined αBB and TAD constrained approach, a comparison was made to an independent TAD method (DYANA (Güntert et al., 1997)) for solving distance restraint prob-

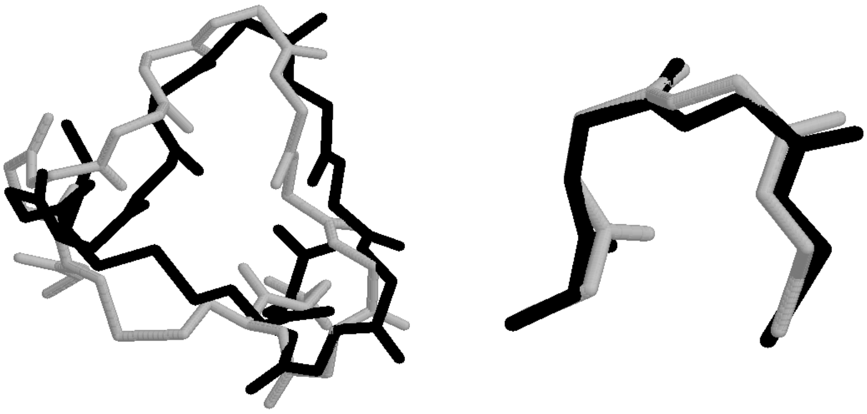

Figure 1.5 Superposition of global minimum (in black) and $\overline{Compstatin}^{Local}$ (in light grey) structures. The left panel shows the full (backbone atom) structure, while the right panel compares only the β-turn region.

lems. 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.

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. 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 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 (Brünger, 1992). However, constrained global optimization was able to identify conformers with significantly lower energies than those obtained from either local minimization or independent TAD algorithms. In particular, the coupled αBB-TAD implementation consistently produced dense populations of low energy conformers.

6. CONCLUSIONS

The importance of the protein folding is evidenced by the large amount of experimental and theoretical research conducted in these areas. Although experimental studies of protein systems are necessary and insightful, the ability to computationally predict and understand the folding of proteins would greatly aid the advancement of the biological and chemical sciences. We have shown that both molecular modeling and global optimization are the dominant factors in the overall equation that will eventually provide a solution to these problems.

In particular, this chapter has focused on the use of ab-initio models, which give rise to a series of complex mathematical problems. The essential component has been the application of deterministic global optimization, namely the αBB algorithm, for solving the resulting problems. Many issues related to the modeling of protein folding have been analyzed and discussed. These observations have highlighted the extreme difficulty of these problems and the crucial interdependence of ab initio modeling and deterministic global optimization approaches.

Acknowledgments

The authors gratefully acknowledge financial support from the National Science Foundation and the National Institutes of Health (R01 GM52032, 1 F32 GM20007).

References

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

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

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

Adjiman, C. S., Dallwig, S., Floudas, C. A., and Neumaier, A. (1998b). A global optimization method for general twice-differentiable nlps - i. theoretical advances. *Comput. Chem. Eng.*, 22:1137–1158.

Adjiman, C. S. and Floudas, C. A. (1996). Rigorous convex underestimators for general twice–differentiable problems. *J. Glob. Opt.*, 9:23–40.

Al-Khayyal, F. A. and Falk, J. E. (1983). Jointly constrained biconvex programming. *Maths Ops Res.*, 8:273–286.

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

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

Anfinsen, C. B., Haber, E., Sela, M., and Jr., F. H. W. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. USA*, 47:1309–1314.

Augspurger, J. D. and Scheraga, H. A. (1996). An efficient, differentiable hydration potential for peptides and proteins. *J. Comp. Chem*, 17:1549–1558.

Becker, O. M. and Karplus, M. (1997). The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.*, 106(4):1495–1517.

Brünger, A. (1992). X-PLOR, version 3.1 a system for x-ray crystallography and nmr. Yale University Press, New Haven, USA.

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

Church, B. W., Orešič, M., and Shalloway, D. (1996). 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.

Czerminski, R. and Elber, R. (1990). Reaction path study of conformational transitions in flexible systems: Applications to peptides. *J. Chem. Phys.*, 92(9):5580–5601.

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

Dejaegere, A. and Karplus, M. (1996). Analysis of coupling schemes in free energy simulations: A unified description of nonbonded contributions to solvation free energies. *J. Phys. Chem.*, 100:11148–11164.

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

Flory, P. J. (1974). Foundations of rotational isomeric state theory and general methods for generating configurational averages. *Macromolecules*, 7(3):381–392.

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

Gerschgorin, S. (1931). Über die abgrenzung der eigenwerte einer matrix. *Izv. Akad. Nauk SSSR, Ser. fiz.-mat.*, 6:749–754.

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

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

Go, N. and Scheraga, H. A. (1976). On the use of classical statistical mechanics in the treatment of polymer chain conformations. *Macromolecules*, 9(4):535–542.

Goldberg, D. (1989). *Genetic algorithms in search, optimization and machine learning.* Addison-Wesley.

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

Honig, B., Sharp, K., and Yang, A. (1993). Macroscopic models of aqueous solutions: Biological and chemical applications. *J. Phys. Chem.*, 97:1101–1109.

Horst, R. and Pardalos, P. M., editors (1995). *Handbook of Global Optimization.* Kluwer Academic Publishers.

Horst, R. and Tuy, H. (1993). *Global optimization: deterministic approaches.* Springer-Verlag, Berlin. 2nd. rev. edition.

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

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

Kang, Y. K., Némethy, G., and Scheraga, H. A. (1987a). Free energies of hydration of solute molecules 2. application of the hydration shell model to nonionic organic molecules. *J. Phys. Chem.*, 91:4109.

Kang, Y. K., Némethy, G., and Scheraga, H. A. (1987b). Free energies of hydration of solute molecules 3. application of the hydration shell model to charged organic molecules. *J. Phys. Chem.*, 91:4118.

Kim, P. S. and Baldwin, R. L. (1990). Intermediates in the folding reactions of small proteins. *Annu. Rev. Biochem.*, 59:631–660.

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

Kitao, A., Hirata, F., and Go, N. (1993). 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.

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

Klepeis, J. L. and Floudas, C. A. (1999). Comparative study of global minimum energy conformations of hydated peptides. *J. Computational Chemistry*, 20:636.

Klepeis, J. L. and Floudas, C. A. (2000). Deterministic global optimization and torsion angle dynamics for molecular structure prediction. *Comp. Chem. Eng.*, 24:1761–1766.

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

Kollman, P. (1993). Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.*, 93:2395–2417.

Leopold, P., Montal, M., and Onuchic, J. (1992). Protein folding funnels : A kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci. USA*, 89:8721–8725.

Levinthal, C. (1968). Are there pathways to protein folding ? *J. Chem. Phys.*, 65:44–45.

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

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

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

Maranas, C. D. and Floudas, C. A. (1995). Finding all solutions of nonlinearly constrained systems of equations. *Journal of Global Optimization*, 7(2):143–182.

McCormick, G. P. (1976). Computability of global solutions to factorable nonconvex programs : Part i – convex underestimating problems. *Math. Programming*, 10:147–175.

Meirovitch, H. and Meirovitch, E. (1997). 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.

Meirovitch, H. and Vásquez, M. (1997). 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.

Morikis, D., Assa-Munt, N., Sahu, A., and Lambris, J. D. (1998). Solution structure of compstatin, a potent complement inhibitor. *Protein Sci.*, 7:619–627.

Némethy, G., Gibson, K. D., Palmer, K. A., Yoon, C. N., Paterlini, G., Zagari, A., Rumsey, S., and Scheraga, H. A. (1992). Energy parameters in polypeptides. 10. *J. Phys. Chem.*, 96:6472–6484.

Neumaier, A. (1990). *Interval Methods for Systems of Equations.* Encyclopedia of Mathematics and its Applications. Cambridge University Press.

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

Perrot, G., Cheng, B., Gibson, K. D., J. Vila, K. A. P., Nayeem, A., Maigret, B., and Scheraga, H. A. (1992). Mseed: A program for the rapid analytical determination of accessible surface areas and their derivatives. *J. Comp. Chem*, 13:1–11.

Ratschek, H. and Rokne, J. (1988). *Computer Methods for the Range of Functions*. Ellis Horwood Series in Mathematics and its Applications. Halsted Press.

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

Sahu, A., Kay, B., and Lambris, J. (1996). Inhibition of human complement by a c3-binding peptide isolated from a phage-displayed random peptide library. *J. Immunol.*, 157:884–891.

Scheraga, H. (1996). *PACK: Programs for Packing Polypeptide Chains*. online documentation.

Schiffer, C. A., Caldwell, J. W., Kollman, P. A., and Stroud, R. M. (1993). Protein structure prediction with a combined solvation free energy-molecular mechanics force field. *Mol. Sim.*, 10:121.

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

Stillinger, F. H. and Weber, T. A. (1984). *J. Chem. Phys.*, 80:4434.

Straatsma, T. P. and McCammon, J. A. (1992). Computational alchemy. *Annu. Rev. Phys. Chem.*, 43:407–435.

Vásquez, M., Némethy, G., and Scheraga, H. A. (1994). Conformational energy calculations on polypeptides and proteins. *Chemical Reviews*, 94:2183–2239.

Šali, A., Shaknovich, E., and Karplus, M. (1996). Thermodynamics and kinetics of protein folding. In *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, volume 23, pages 199–213. American Mathematical Society.

Wesson, L. and Eisenberg, D. (1992). Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Science*, 1:227.

Williams, R. L., Vila, J., Perrot, G., and Scheraga, H. A. (1992). Empirical solvation models in the context of conformational energy searches: Application to bovine pancreatic trypsin inhibitor. *Proteins*, 14:110–119.