Predicting Peptide Structures using NMR Data and Deterministic Global Optimization

J. L. Klepeis, C.A. Floudas,*

Department of Chemical Engineering

Princeton University

Princeton, N.J. 08544-5263

D. Morikis,

Department of Chemistry and Biochemistry

University of California San Diego

La Jolla, CA 92093-0359

and

J.D. Lambris

Department of Pathology and Laboratory Medicine

University of Pennsylvania

Philadelphia, PA 19104

*Author to whom all correspondence should be addressed; Tel.: (609) 258-4595; Fax: (609) 258-0211; email: floudas@titan.princeton.edu

Abstract

The ability to analyze large molecular structures by NMR techniques requires efficient methods for structure calculation. Currently there are several widely available methods for tackling these problems, which, in general, rely on the optimization of penalty-type target functions in order to satisfy the conformational restraints. Typically, these methods combine simulated annealing protocols with molecular dynamics and local minimization, either in distance or torsional angle space. In this work, both a novel formulation and algorithmic procedure for the solution of the NMR structure prediction problem is outlined. First, the unconstrained, penalty-type structure prediction problem is reformulated using nonlinear constraints, which can be individually enumerated for all, or subsets, of the distance restraints. In this way, the violation can be controlled as a constraint, in contrast to the usual penalty-type restraints. In addition, the customary simplified objective function is replaced by a full atom force field in the torsional angle space. This guarantees a better description of atomic interactions, which dictate the native structure of the molecule along with the distance restraints. The second novel portion of this work involves the solution method. Rather than pursue the typical simulated annealing procedure, this work relies on a deterministic method, which theoretically guarantees that the global solution can be located. This branch and bound technique, based on the αBB algorithm, has already been successfully applied to the identification of global minimum energy structures of peptides modeled by full atom force fields. Finally, the approach is applied to the Compstatin structure prediction and it is found to possess some important merits when compared to existing techniques.

Keywords: Global optimization, Structure calculation, NMR, Compstatin, Complement inhibitor

1 Introduction

The use of nuclear magnetic resonance (NMR) data has become a widely developed technique for determining protein structures. The data obtained from NMR studies consist of distance and 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 torsional angles are based on the measurement of coupling constants and analysis of proton chemical shifts. Together, this information can be used to formulate a nonlinear optimization problem, whose solution should provide the correct protein structure.

However, the structure prediction problem is extremely complex for several reasons. The major difficulty is the imprecision of distance information due to the influence of spin diffusion and internal dynamics on the relationship between the NOE intensity and the interproton distance. Even if this distance information is consistent, the number of distance limits is generally much smaller than needed to determine a unique structure. Therefore, a simple distance geometry approach is not sufficient.

In order to address these problems, the structure prediction problem is transformed to an optimization problem based on a hybrid energy function of the following form:

$$E = E_{forcefield} + W_{nmr} E_{nmr}. \tag{1}$$

The energy, E, specified by this target function now includes a chemical description of the protein conformation through the use of an empirical force field, $E_{forcefield}$. However, these force field potentials are generally much simpler representations of typical all atom force fields. The distance and dihedral angle restraints are included (as in pure distance geometry problems) in the objective function, although they now appear as weighted (with weight W_{nmr}) penalty terms that should be driven to zero. Both terms are complicated functions of the atomic coordinates, and this problem has generally been referred to as the multiple-minima problem. That is, the prediction of the global minimum energy structure, which should correspond to the correct structure, is hindered by the presence of many local energy minima with relatively high energy barriers.

Calculating three-dimensional structures using NMR data is therefore dependent on the development of efficient optimization methods. Typically, one of two optimization methods have been employed. The first is based on the minimization of a variable target function of distance restraints and nonbonded contacts in torsional angle space[1,2,3]. The second relies on optimization of a hybrid energy function by coupling simulated annealing with molecular dynamics in Cartesian coordinate space[2,4]. For large proteins, these methods require relatively long computation times and generally provide a low yield of acceptable conformations. This is mainly a result of the multiple-minima of the objective function, and the difficulty of escaping local minima using molecular dynamics in Cartesian space. More recent methods have implemented torsion angle dynamics (TAD) and have been shown to be more effective than Cartesian coordinate dynamics[5,6]. 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. An overview of available methods for predicting three-dimensional protein structures can be found elsewhere[7,8,9].

In this work, a novel formulation and global optimization approach are proposed for the three-dimensional structure prediction problem using NMR data. The proposed method is based on a *constrained* formulation, which differs from the traditional formulations that employ penalty function methods. In addition, the nonlinear objective function is represented by a detailed full atom force field, rather than simplified nonbonded potential terms. The solution of this novel NMR based formulation is accomplished by developing an algorithm based on the ideas of the αBB deterministic global optimization approach[10,11,12,13,14]. In the next two sections the development of this novel method is presented, which is then followed by a detailed case study for the prediction of the three-dimensional structure of the 13-residue synthetic peptide, Compstatin[15].

2 Theory

2.1 Energy Modeling

The target function shown in Equation (1) can rewritten in the following form:

$$E = E_{bonds} + E_{angles} + E_{chiral,planar} + E_{distance} + E_{dihedral} + E_{forcefield}. \tag{2}$$

In this equation, the E_{nmr} term is expanded and the weighting factor (W_{nmr}) is incorporated separately into each individual term. The first three terms (E_{bonds}, E_{angles} and $E_{chiral,planar}$) are typically treated as quadratic harmonic potentials for bond lengths, bond angles and chirality and planarity. For example :

$$E_{bonds} = \sum_{bonds} k_r (r - r_o)^2, \tag{3}$$

$$E_{angles} = \sum_{angles} k_\theta (\theta - \theta_o)^2, \tag{4}$$

$$E_{chiral,planar} = \sum_{chiral,planar} k_\phi (\phi - \phi_o)^2. \tag{5}$$

Here r_o, θ_o and ϕ_o represent reference values for the bond lengths, angles and dihedral angles, respectively. The k_r, k_θ and k_ϕ are the corresponding force constants.

The fourth term of Equation (2) accounts for the objective function contribution corresponding to experimental distance restraints. This function can take several forms, although the most general form corresponds to a simple square well potential, which includes a summation over both upper and lower distance violations (i.e., $E_{distance} = E^{upper}_{distance} + E^{lower}_{distance}$). When considering upper distance restraints this becomes:

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

The squared violation energy is considered only when the calculated distance d_j exceeds the upper reference distance d_j^{upper}. This squared violation is 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}:

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

It should be noted that, in general, penalty terms enforcing both upper and lower distance bounds are used. In this case the condition $d_j^{lower} \le d_j \le d_j^{upper}$ must be enforced.

When considering dihedral angle restraints, represented by term 5 in Equation (2), a form similar to Equations (6) and (7) is often used. The total violation, $E_{dihedral}$, is a sum over upper and lower violations (i.e., $E_{dihedral} = E_{dihedral}^{upper} + E_{dihedral}^{lower}$). 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 $[\omega_j^{lower} - \Delta HW_{\omega_j}, \omega_j^{upper} + \Delta HW_{\omega_j}]$, where ΔHW_{ω_j} is equal to half the excluded range of dihedral angle values (i.e., $\Delta HW_{\omega_j} = pi - (\omega_j^{upper} - \omega_j^{lower})/2$). This results in the following equations:

$$E_{dihedral}^{upper} = \sum_{upper} \begin{cases} A_j\left(1 - 2\left[\frac{\omega_j - \omega_j^{upper}}{2\pi - (\omega_j^{upper} - \omega_j^{lower})}\right]^2\right)(\omega_j - \omega_j^{upper})^2 & \text{if } \omega_j > \omega_j^{upper}, \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

$$E_{dihedral}^{lower} = \sum_{lower} \begin{cases} A_j\left(1 - 2\left[\frac{\omega_j - \omega_j^{lower}}{2\pi - (\omega_j^{upper} - \omega_j^{lower})}\right]^2\right)(\omega_j - \omega_j^{lower})^2 & \text{if } \omega_j < \omega_j^{lower}, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Finally, the last term in Equation (2) refers to the force field energy expressions used to model the nonbonded interactions of the protein. These often correspond to simple Van der Waals repulsion terms. More detailed force fields employ 6-12 Lennard-Jones and modified 10-12 Lennard-Jones terms to model nonbonded and hydrogen bonded interactions, respectively. An additional Coulombic electrostatic term may also be included.

In practice, when considering NMR restraints, force field terms are often simplified to include only simple geometric energy terms, such as quartic Van der Waals repulsions. Such objective functions neglect rigorous modeling of energetic terms in order to ensure that experimental distance violations are minimized. In fact, a simple representation for the objective function using torsional angle dynamics would be :

$$E_{simple} = E_{distance} + E_{dihedral}. \tag{10}$$

In this case, the target function does not include bond, angle or chirality/planarity violation energies. Notice that when all restraints are satisfied, the objective function is driven to zero.

In this work, a more detailed modeling approach is proposed by using the ECEPP/3 force field[16]. For 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. That is, E_{bonds}, E_{angles} and $E_{chiral,planar}$ are inherently equal to zero. The total force field energy, $E_{forcefield}$, is calculated as the sum of the electrostatic, nonbonded, hydrogen bonded, and torsional contributions. 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, while the development of the appropriate parameters is discussed and reported elsewhere[16].

When considering a simple unconstrained minimization, this approach corresponds to an objective function defined by:

$$E_{detailed} = E_{distance} + E_{dihedral} + E_{ECEPP/3}. \tag{11}$$

This formulation is similar to Equation (10) in that the E_{bonds}, E_{angles} and $E_{chiral,planar}$ can be neglected. However, the detailed energy modeling greatly increases the computational complexity of the objective function. It should also be noted that although distances correspond to quadratic terms in Cartesian coordinate space, their transformation to internal coordinate space results in complex, highly nonlinear functions. That is there is not a one-to-one correspondence between distances and internal coordinates. The advantage for working in dihedral angle space is that the variable set decreases, with the disadvantage being the increased nonconvexity of the energy hypersurface.

2.2 Global Optimization

The determination of a three dimensional protein structure defines an optimization problem for which the objective function is defined by the target functions outlined in the previous section. Methods for addressing this optimization problem are outlined in the following sections. The first section presents the standard penalty function approach used in structure determination problems. This is followed by a section describing a deterministic global optimization method for solving these (unconstrained) problems. This approach, based on the αBB branch and bound algorithm[10,11,12,13,14], can deterministically locate global minimum energy structures without the bias of initial structure selection usually associated with stochastic searches. The third section introduces a novel *constrained* formulation which is easily incorporated within with the αBB global optimization procedure, while the final section provides some details regarding this constrained formulation.

2.2.1 Penalty Function Formulation

A standard procedure for addressing the global optimization problem involving NMR and dihedral angle restraints consists of a combination of discrete geometry (metric method) optimization using a simulated annealing protocol coupled with molecular or torsional angle dynamics. 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 order to avoid trapping in local minimum energy wells.

However, the minimization of complex target functions necessitates the use of rigorous global optimization approaches. In this work, a detailed force field potential is employed in the context of a conformational energy search using NMR restraints. This typical penalty type formulation (for distance restraints) can be written as:

$$\min \quad E_{detailed}(\phi_i, \psi_i, \omega_i, \chi_i^k, \theta_j^N, \theta_j^C) \quad = \quad E_{distance} + E_{dihedral} + E_{ECEPP/3}, \tag{12}$$

$$\begin{aligned}
\text{subject to} \quad \phi_i^L &\leq \phi_i \leq \phi_i^U, \quad i = 1, \dots, N_{RES}, \\
\psi_i^L &\leq \psi_i \leq \psi_i^U, \quad i = 1, \dots, N_{RES}, \\
\omega_i^L &\leq \omega_i \leq \omega_i^U, \quad i = 1, \dots, N_{RES}, \\
\chi_i^{k,L} &\leq \chi_i^k \leq \chi_i^{k,U}, \quad i = 1, \dots, N_{RES}, \; k = 1, \dots, K^i, \\
\theta_j^{N,L} &\leq \theta_j^N \leq \theta_j^{N,U}, \quad j = 1, \dots, J^N, \\
\theta_j^{C,L} &\leq \theta_j^C \leq \theta_j^{C,U}, \quad j = 1, \dots, J^C.
\end{aligned}$$

Here $i = 1, \dots, N_{RES}$ is an indexed set describing the sequence of amino acid residues in the peptide chain. There are $\phi_i, \psi_i, \omega_i, \; i = 1, \dots, N_{RES}$ dihedral angles along the backbone of this peptide. In addition, K^i denotes 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. Also, $\phi_i^L, \psi_i^L, \omega_i^L, \chi_i^{k,L}, \theta_j^{N,L}, \theta_j^{C,L}$ and $\phi_i^U, \psi_i^U, \omega_i^U, \chi_i^{k,U}, \theta_j^{N,U}, \theta_j^{C,U}$ represent lower and upper bounds on the dihedral angles $\phi_i, \psi_i, \omega_i, \chi_i^k, \theta_j^N, \theta_j^C$. In the simplest case, the energy function corresponds to a target function of the form given in Equation (10). However, in this work, $E_{detailed}$ includes both a complex force field modeled by ECEPP/3, and NMR distance and dihedral angle restraints, as shown in Equation (11). The solution of (12) using either objective function constitutes an unconstrained global minimization problem. A deterministic method for solving such problems is given in the next section. As will be shown, a novel reformulation can also be used in order to effectively treat this problem as a *constrained* global conformational energy search.

2.2.2 αBB Deterministic Global Optimization

When NMR restraints are not considered, the formulation given by Equation (12) corresponds to the traditional protein folding problem[17]. That is, the problem involves the global minimization of conformational energy with respect to the independent dihedral angles. Typically, the lower and upper bounds for these variables are set to -π and π, respectively. In this case, a detailed atomistic-level energy produces a multidimensional surface with an astronomically large number of local minima. Because the objective function has many local minima, using local optimization techniques necessarily depends on the initial points selected. Therefore, rigorous global optimization algorithms are needed to effectively locate the global minimum corresponding to the native state of the protein. A large number of techniques have been developed to search this nonconvex conformational space. In general, the major limitation is that these methods also depend strongly on the supplied initial conformations. As a result, there is no guarantee for global convergence because large sections of the domain space may be bypassed. In order to overcome these difficulties, the αBB global optimization approach[10,11,12,13,14] has been extended to identifying global minimum energy conformations of peptides. The development of this branch and bound method was motivated by the need for an algorithm that could guarantee convergence to the global minimum of nonlinear optimization problems with twice–differentiable functions[18]. The application of this algorithm to the minimization of potential energy functions was first introduced for microclusters[19,20], and small acyclic molecules[21,22]. The αBB approach has also been applied to general constrained optimization problems[10,12,13,14]. In more recent work, the algorithm has been shown to be successful for isolated peptide systems using the realistic ECEPP/3 potential energy model[23,24], and including several solvation effects[25,26].

The αBB global optimization algorithm effectively brackets the global minimum solution by developing converging sequences of lower and upper bounds. These bounds are refined by iteratively partitioning the initial domain. Upper bounds on the global minimum are obtained by local minimizations of the original energy function, E. Lower bounds belong to the set of solutions of the convex lower bounding functions, which are constructed by augmenting E with the addition of separable quadratic terms. The lower bounding functions, L, of the energy hypersurface can be expressed in the following manner:

$$
\begin{aligned}
L \;=\; E + \quad & \sum_{i=1}^{N_{RES}} \alpha_{\phi_i} \left(\phi_i^L - \phi_i\right) \left(\phi_i^U - \phi_i\right) + \\
& \sum_{i=1}^{N_{RES}} \alpha_{\psi_i} \left(\psi_i^L - \psi_i\right) \left(\psi_i^U - \psi_i\right) + \\
& \sum_{i=1}^{N_{RES}} \alpha_{\omega_i} \left(\omega_i^L - \omega_i\right) \left(\omega_i^U - \omega_i\right) + \\
& \sum_{i=1}^{N_{RES}} \sum_{k=1}^{K^i} \alpha_{\chi_i^k} \left(\chi_i^{k,L} - \chi_i^k\right) \left(\chi_i^{k,U} - \chi_i^k\right) + \\
& \sum_{j=1}^{J^N} \alpha_{\theta_j^N} \left(\theta_j^{N,L} - \theta_j^N\right) \left(\theta_j^{N,U} - \theta_j^N\right) + \\
& \sum_{j=1}^{J^C} \alpha_{\theta_j^C} \left(\theta_j^{C,L} - \theta_j^C\right) \left(\theta_j^{C,U} - \theta_j^C\right) .
\end{aligned}
\tag{13}
$$

Again, $\phi_i^L, \psi_i^L, \omega_i^L, \chi_i^{k,L}, \theta_j^{N,L}, \theta_j^{C,L}$ and $\phi_i^U, \psi_i^U, \omega_i^U, \chi_i^{k,U}, \theta_j^{N,U}, \theta_j^{C,U}$ represent lower and upper bounds on the dihedral angles $\phi_i, \psi_i, \omega_i, \chi_i^k, \theta_j^N, \theta_j^C$. The α parameters represent nonnegative parameters which must be greater or equal to the negative one–half of the minimum eigenvalue of the Hessian of E over the defined domain. Rigorous bounds on the α parameters can be obtained through a variety of approaches[27,12,13,21]. The overall effect of these terms is to overpower the nonconvexities of the original nonconvex terms by adding the value of 2α to the eigenvalues of the Hessian of E.

Once solutions for the upper and lower bounding problems have been established, the next step is to modify these problems for the next iteration. This is accomplished by successively partitioning the initial domain into smaller subdomains. For the protein conformation problems, it has been found that an effective partitioning strategy involves bisecting the same variable dimension across all nodes at a given level. In order to ensure non–decreasing lower bounds, the hyper–rectangle to be bisected is chosen by selecting the region which contains the infimum of the minima of lower bounds. A non–increasing sequence for the upper bound is found by solving the nonconvex problem, E, locally and selecting it to be the minimum over all the previously recorded upper bounds. Obviously, if the single minimum of L for any hyper–rectangle is greater than the current upper bound, this hyper–rectangle can be discarded because the global minimum cannot be within this subdomain (fathoming step). The computational requirement of the αBB algorithm depends on the number of variables (global) on which branching occurs. Therefore, these global variables need to be chosen carefully.

An important implication of the αBB branch and bound approach is the implicit treatment of dihedral angle restraints. Specifically, since partitioning of the dihedral angle space represents an inherent part of the αBB problem formulation, these bounds can be easily satisfied by defining appropriate upper and lower bounds on these variables. Therefore, the dihedral angle restraint energy, $E_{dihedral}$, of the target function given in Equation (11) is always driven to zero.

2.2.3 Novel Constrained Formulation

Since the αBB approach implicitly handles dihedral angle restraints, the objective function given in (12) only effectively includes the forcefield, $E_{ECEPP/3}$, and distance restraint, $E_{distance}$, energies. The objective function can also be reformulated by treating distance restraints as a set of general nonlinear constraint(s); that is, the distance restraint energy, $E_{distance}$, is not a required part of the objective function. This *constrained* formulation becomes:

$$
\begin{aligned}
\min \quad & E_{forcefield} \quad (\phi_i, \psi_i, \omega_i, \chi_i^k, \theta_j^N, \theta_j^C), \qquad (14)\\
\text{subject to} \quad & E_l^{distance} \quad (\phi_i, \psi_i, \omega_i, \chi_i^k, \theta_j^N, \theta_j^C) \ \leq \ E_l^{ref} \quad l = 1, \ldots, N_{CON},\\
& \phi_i^L \leq \quad \phi_i \ \leq \ \phi_i^U, \ \ i = 1, \ldots, N_{RES},\\
& \psi_i^L \leq \quad \psi_i \ \leq \ \psi_i^U, \ \ i = 1, \ldots, N_{RES},\\
& \omega_i^L \leq \quad \omega_i \ \leq \ \omega_i^U, \ \ i = 1, \ldots, N_{RES},\\
& \chi_i^{k,L} \leq \quad \chi_i^k \ \leq \ \chi_i^{k,U}, \ \ i = 1, \ldots, N_{RES}, \ k = 1, \ldots, K^i,\\
& \theta_j^{N,L} \leq \quad \theta_j^N \ \leq \ \theta_j^{N,U}, \ \ j = 1, \ldots, J^N,\\
& \theta_j^{C,L} \leq \quad \theta_j^C \ \leq \ \theta_j^{C,U}, \ \ j = 1, \ldots, J^C.
\end{aligned}
$$

The lower and upper variable bounds ($\phi_i^L, \psi_i^L, \omega_i^L, \chi_i^{k,L}, \theta_j^{N,L}, \theta_j^{C,L}$ and $\phi_i^U, \psi_i^U, \omega_i^U, \chi_i^{k,U}, \theta_j^{N,U}, \theta_j^{C,U}$) are first modified to correspond to upper and lower dihedral angle restraints, rather than the entire domain of -π to π. Distance restraints, $E_l^{distance}$, are now constrained to be below a total reference energy, E_l^{ref}. These distance constraints are typically identical in form to the summation of the square well violation energies, as given in Equations (6) and (7), although other functional forms may be used. In addition, note that formulation (14) may include a full enumeration of all distance restraints or a selected subset of these restraints. In general, the constrained formulation is more rigorous than a penalty function approach because the choice of E_l^{ref} strictly determines the extent to which each set of restraints must be satisfied. In addition, in the limit that E_l^{ref} approaches zero, all restraints are implicitly enforced. Since this violation energy is imposed as a set of constraints, all local solutions are also required to meet this specification. As a result, the proposed constrained formulation has the advantage of not requiring the specification of penalty coefficients, which are typically updated through variable target function methods fashion when using the unconstrained penalty-type approaches.

2.2.4 Lower Bounding via the αBB

In order to treat the NMR structure prediction problem via the constrained formulation, a number of modifications must be made to the αBB methodology previously outlined. In particular, the identification of valid lower bounds on the global solution of the nonconvex problem relies on the fact that the underestimating problem generated in each subdomain must be convex. The development of the appropriate convex lower bounding function for the objective function has already been discussed. In addition, all inequality constraints in the lower bounding problem must be convex, which implies that all inequality constraints appearing in Equation (14) must be replaced by their convex relaxation. For each constraint the following expression is used :

$$
\begin{aligned}
L_l^{distance} = E_l^{distance} + & \sum_{i=1}^{N_{RES}} \alpha_{\phi_i,l}^{distance} \left(\phi_i^L - \phi_i\right) \left(\phi_i^U - \phi_i\right) + \\
& \sum_{i=1}^{N_{RES}} \alpha_{\psi_i,l}^{distance} \left(\psi_i^L - \psi_i\right) \left(\psi_i^U - \psi_i\right) + \\
& \sum_{i=1}^{N_{RES}} \alpha_{\omega_i,l}^{distance} \left(\omega_i^L - \omega_i\right) \left(\omega_i^U - \omega_i\right) + \\
& \sum_{i=1}^{N_{RES}} \sum_{k=1}^{K^i} \alpha_{\chi_i^k,l}^{distance} \left(\chi_i^{k,L} - \chi_i^k\right) \left(\chi_i^{k,U} - \chi_i^k\right) + \\
& \sum_{j=1}^{J^N} \alpha_{\theta_j^N,l}^{distance} \left(\theta_j^{N,L} - \theta_j^N\right) \left(\theta_j^{N,U} - \theta_j^N\right) + \\
& \sum_{j=1}^{J^C} \alpha_{\theta_j^C,l}^{distance} \left(\theta_j^{C,L} - \theta_j^C\right) \left(\theta_j^{C,U} - \theta_j^C\right).
\end{aligned}
\tag{15}
$$

The $\alpha_l^{distance}$ represent nonnegative parameters which must be greater or equal to the negative one–half of the minimum eigenvalue of the Hessian of $E_l^{distance}$ over the defined domain. These functions must be developed for each constraint belonging to the set $l = 1, \ldots, N_{CON}$. Rigorous bounds on these α parameters can be obtained via several methods[27,12,13,21].

Therefore, the full lower bounding formulation for the constrained NMR problem can be expressed as :

$$
\min \quad L_{forcefield} \quad (\phi_i, \psi_i, \omega_i, \chi_i^k, \theta_j^N, \theta_j^C), \tag{16}
$$

$$
\begin{array}{llcll}
\text{subject to} & L_l^{distance} & (\phi_i, \psi_i, \omega_i, \chi_i^k, \theta_j^N, \theta_j^C) & \leq & E_l^{ref} \quad l = 1, \ldots, N_{CON}, \\
& \phi_i^L \leq & \phi_i & \leq & \phi_i^U, \ \ i = 1, \ldots, N_{RES}, \\
& \psi_i^L \leq & \psi_i & \leq & \psi_i^U, \ \ i = 1, \ldots, N_{RES}, \\
& \omega_i^L \leq & \omega_i & \leq & \omega_i^U, \ \ i = 1, \ldots, N_{RES}, \\
& \chi_i^{k,L} \leq & \chi_i^k & \leq & \chi_i^{k,U}, \ \ i = 1, \ldots, N_{RES}, \ k = 1, \ldots, K^i, \\
& \theta_j^{N,L} \leq & \theta_j^N & \leq & \theta_j^{N,U}, \ \ j = 1, \ldots, J^N, \\
& \theta_j^{C,L} \leq & \theta_j^C & \leq & \theta_j^{C,U}, \ \ j = 1, \ldots, J^C,
\end{array}
$$

In this formulation, variable bounds are specific to the subdomain for which the lower bounding functions are constructed. $L_{forcefield}$ refers to the convex representation of the objective function (Equation (13)), while $L_l^{distance}$ denotes the convex relaxation of the inequality constraints as given in Equation (15). As before, a converging sequence of upper and lower bounding values are developed, although these values now depend on the solution of the problems given by Equations (14) and (16), respectively.

3 Algorithmic Steps

A description of the steps involved in the solution of the NMR structure prediction problem using the constrained αBB approach can be generalized to any force field model and any routine for locally solving constrained optimization problems. In this work, the αBB approach is interfaced with PACK[28] and NPSOL[29]. PACK is used to transform to and from Cartesian and internal coordinate systems, which is needed to obtain function and gradient contributions for the ECEPP/3 force field and the distance constraint equations. NPSOL is a local nonlinear optimization solver that is used to locally solve the constrained upper and lower bounding problems in each subdomain.

The implementation can be broken down into two main phases : initialization and computation. The basic steps of the initialization phase are as follows :

(1) Choose the set of global variables. Since the bounds on these variables will be refined during the course of global optimization, they should be selected based on their overall effect on the structure of the molecule. In this work (and in general) the ϕ and ψ dihedral angles provide the largest structural variability, and are chosen to constitute the global variable set.

(2) Set upper and lower bounds on all dihedral angles (variables). If information is not available for a given dihedral angle, the variable bounds are set to [-π,π]. Since a constrained local optimization solver is used, these bounds are strictly enforced.

(3) Identify the set of NOE derived distance restraints to be used in the constraints. In general, this set can include all intra- and inter-residue restraints. In this work, only backbone sequential and medium/long range information was used in developing the constraints, because intra-residue restraints are less likely to affect the overall fold. In addition, although multiple constraints with varying weights can be handled, all distance information was formulated as one constraint ($N_{CON} = 1$) with constant weighting for simplicity.

(4) Choose the value of E_l^{ref} to be used in constraints. This can be determined by simply performing several local constrained optimizations or possibly a short global optimization run with simplified energy models. In this work, information based on X-PLOR[4] results was used to define the E^{ref} parameter (see below).

(5) Identify initial α values for both the objective function and constraints, as defined in sections 2.2.2 and 2.2.4, respectively.

(6) Set initial best upper bound to an arbitrarily large value.

The computation phase of the algorithm involves an iterative approach, which depends on the refinement of the original domain by partitioning along the global variables. In each subdomain, upper and lower bounding problems based on the formulations given in (14) and (16), respectively, are solved locally and used to develop the sequence of converging upper and lower bounds. The basic steps are as follows:

(1) The original domain (defined above) is partitioned along one of the global variables.

(2) Lower bounding functions for both the objective (Equation (13)) and constraints (Equation (15)) are constructed in both subdomains. A constrained local minimization (with NPSOL) is performed using the following procedure :

  (A) 100 random points are generated and used for evaluation of the lower bounding objective function and constraints.

  (B) The point with the minimum objective function value is used as a starting point for local minimization (of formulation (16)) using NPSOL.

  (C) If the minimum value found is greater than the current best upper bound the subdomain can be fathomed (global minimum is outside region), otherwise the solution is stored.

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

  (A) 100 random points are generated and used for evaluation of the objective function and constraints.

  (B) The point with the minimum objective function value and feasible constraints is used as a starting point for local minimization (of formulation (14)) using NPSOL. If a feasible starting point is not found, local minimization is not performed.

  (C) All feasible solutions are stored.

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

(5) The subdomain with the current minimum value of $L_{forcefield}$ is selected and partitioned along one of the global variables.

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

The location of the global minimum relies on effectively solving the upper bounding problem locally. In addition, convergence to this global minimum can be enhanced by consistently identifying low energy solutions. Although this property is not required to prove convergence to the global minimum (because subsequent partitioning revisits regions containing the global solution), it can have important practical implications for high dimensional problems. These observations illustrate the need for reliably locating low energy feasible points. For the Compstatin example, the approach outlined above proved to be sufficient; however, this performance may not be expected for all examples. Along these lines, we are developing methods that combine aspects of torsion angle dynamics (TAD) and constrained local minimization, within the framework of the constrained αBB approach[30].

4 Results and Discussion

4.1 Compstatin Case Study : Traditional Solution Structures

Compstatin is a synthetic 13-residue (ICVVQDWGHHRCT) cyclic peptide that binds to C3 (third component of complement) and inhibits complement activation[31]. The synthetic peptide is cyclic, with a disulfide bridge between the Cys^2 and Cys^{12} residues. The solution structure was previously identified using two-dimensional NMR techniques[15]. A total of 30 backbone sequential (including H^β – backbone), 23 medium and long range (including disulfide) and 82 intra-residue NOE restraints were identified. In addition, 7 ϕ angle and 2 χ_1 angle dihedral restraints were provided. In previous work[15], a traditional distance geometry–simulated annealing protocol was utilized to minimize the associated target function (as in Equation (2)) in the Cartesian coordinate space using the program X-PLOR[4]. This target function consisted of quadratic harmonic potential terms for bonds, angle, planarity and chirality. The force field energy, $E_{forcefield}$, was simplified to account for only quartic Van der Waals repulsion of non–bonded contacts. That is, no hydrogen bonding, electrostatic or Lennard-Jones type empirical potential energy terms were included. NOE distance and dihedral angle restraints were modeled using a quadratic square well potential, similar to those of Equations (6) and (7). In addition $^3J_{NH-H\alpha}$ coupling constant restraints were included as harmonic potentials[15].

Employing typical NMR refinement protocols resulted in a family of structures with similar geometries in the Gln^5–Gly^8 region. Using an ensemble of 21 refined structures, an average 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. This information is displayed in Table I.

4.2 Local Minimization

The consistency of the ensemble of Compstatin solution structures was determined by evaluating distance restraints for each of the original 21 structures (accession number 1a1p at the Brookhaven Protein Data Bank, http://www.pdb.bnl.gov), as well as the average Compstatin conformation. In considering distance restraints, only backbone sequential and medium/long range NOE were considered. That is, the 82 intra-residue restraints were neglected since they are less likely to effect the overall fold of the Compstatin peptide. This results in a total of 52 restraints, with an additional restraint on the distance between the sulfur atoms forming the disulfide bridge (a total of 53 distance restraints). In order to quantify these results the sum of distance violations (D_{VIO}) and a violation energy (E_{VIO}) is reported for each of the original PDB structures. The sum of distance violations corresponds to the sum of the absolute values of the upper and lower violations based on Equations (6) and (7). When all restraints are satisfied, this summation goes to zero. The violation energy is calculated by combining Equations (6) and (7):

$$E_{VIO} = E^{upper}_{distance} + E^{lower}_{distance}. \tag{17}$$

In these calculations, the value of the weighting factor (A_j) is assumed to be constant and set equal to 50 kcal/mol/$\mathring{A}^2$. Table II summarizes this information.

The results shown in Table II indicate that the average structure ($\overline{Compstatin}$) possesses the largest value of D_{VIO}, as well as the third largest violation energy. The smallest distance violation and energy is given by structure number 8 ($< Compstatin >_8$). These results provide a range of comparison for total distance violations and violation energies. In addition, the analysis is used to set the value of E^{ref} (from Equation (14)) to 200 kcal/mol. This value is chosen so that the sum of the violation energies will necessarily result in an improvement over the violation energy for the average Compstatin structure, $\overline{Compstatin}$.

In order to measure the performance of the proposed global optimization approach, the ensemble and average Compstatin structures ($< Compstatin >$ and $\overline{Compstatin}$) were used as starting points for local minimization, as defined by (14). Since PACK (and thus ECEPP/3) builds peptide structure with fixed bond lengths and bond angles (in the internal coordinate, rather than Cartesian coordinate space), the corresponding Compstatin PDB structures could only be used to derive dihedral angle values. These dihedral angles were then used as input to directly evaluate the corresponding force field energy. Because the differences in bond lengths and bond angles propagates through the generation of the corresponding ECEPP/3 structure, an inherent RMSD exists between the PDB structure and the ECEPP/3 generated structure. For example, when using the set of dihedral angles calculated from the $\overline{Compstatin}$ PDB, the ECEPP/3 structure possesses a 0.581 $\mathring{A}$ all atom RMSD (all heavy atoms in backbone and side chains) with respect to the original $\overline{Compstatin}$ structure. The corresponding ECEPP/3 energy equals 519.2 kcal/mol. In addition, due to the differences in bond lengths and angles, the distance violation for the ECEPP/3 structure ($\overline{Compstatin_{ECEPP}}$) increases from 6.9 to 8.7 $\mathring{A}$, which results in a subsequent increase in violation energy to 315 kcal/mol. The superposition of the original and ECEPP/3 $\overline{Compstatin}$ conformations is shown in Figure 2.

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

Structural comparisons between these structures were also quantified using RMSD calculations. These results are shown in Tables IV through VII. The first two tables include all atom and backbone RMSD values between the original (PDB) average Compstatin structure ($\overline{Compstatin}$) and the ensemble of 21 original Compstatin PDB structures ($< Compstatin >_i$). When considering all heavy atoms, these values (see column 2, Table IV) are all clustered near a value of 2 $\AA$. When considering only backbone atoms (see column 2, Table V) the range of values generally fall between 1 - 2 $\AA$. The third column (see both Tables IV and V) reports RMSD values between the original PDB structures and their locally minimized counterparts. In general, these values are larger, which indicates a significant conformational change during local minimization. Finally, the fifth column (see both Tables IV and V) provides a comparison similar to that given by the corresponding second columns, except the structures correspond to local minimum rather than original PDB structures. In general, these RMSD values follow the same trends as the second column (see both Tables IV and V), and in some cases the backbone RMSD are smaller than for the original structures. This indicates that local minimization is providing structural differences that are on the same order as those provided by the original structures. The second set of tables (Tables VI and VII) repeats this analysis, but for the reduced sequence involving residues 5 to 8 (the β-turn region). When considering all heavy atoms, the RMSD values are similar, with most values falling within the 0.5 to 1 $\AA$ range. These results indicate that the β-turn is a common structural feature, even when comparing the original PDB structures to their locally minimized counterparts. A similar trend is observed for the backbone atom RMSD values. The effect of local minimization on conserving the β-turn structure is even more apparent when considering the relatively low and consistent RMSD values of the last column, which compares (residues 5 to 8) the local minimum of the average Compstatin structure ($\overline{Compstatin}^{Local}$) to the local minimum of the individual Compstatin structures ($< Compstatin >_i^{Local}$). Plots for the superpositioning (all atom) of $< Compstatin >_{13}^{Local}$ and the average local minimum structure ($\overline{Compstatin}^{Local}$) are given in Figure 3. The superpositioning of these two structures results in two of the smallest RMSD values, as given in Tables IV and VI.

4.3 Global Minimization

A full global minimization of the Compstatin structure was then performed according to the *constrained* implementation outlined in section 3. In total, Compstatin possesses 73 independent torsion angles, of which 26 (all ϕ and ψ) were treated globally, while the remaining were allowed to vary locally. As with the local minimizations, the same set of restraints were used to formulate the nonlinear constraint, with a constant 50 kcal/mol/$\AA$ weighting factor (A_j) and a constraint parameter (E^{ref}) equal to 200 kcal/mol. The lowest energy structure satisfying the distance constraint and dihedral angle bounds provided an ECEPP/3 energy of -85.71 kcal/mol, which is lower in energy than any of the local minimum structures given in Table III. The global minimization required approximately 40 CPU hours on a HP C160. As with the local minimizations, the global minimum structure reached the 200 kcal/mol bound on the violation energy constraint. The total distance violation (D_{VIO}) equaled 6.690 $\AA$, which is near the average distance violation form those local minimum structures given in Table III.

A number of RMSD calculations were performed to further quantify the structural differences between the global minimum energy structure and the other Compstatin structures. These results are given in Tables VIII and IX. Table VIII provides (all atom and backbone atom) RMSD values between the full local minimum energy structures ($< Compstatin >_i^{Local}$ and $\overline{Compstatin}^{Local}$) and the global minimum energy structure. When comparing backbone RMSDs, the $< Compstatin >_9^{Local}$, $< Compstatin >_{21}^{Local}$, $< Compstatin >_{19}^{Local}$ and $< Compstatin >_{17}^{Local}$ structures offer the best correspondence with the global minimum energy structure. These structures also correspond to four of the lowest energy local minimum, as given in Table III. This indicates that some of the lowest energy conformers exhibit similar backbone structural characteristics. However, it is interesting to note that the lowest energy local minimum, $< Compstatin >_2^{Local}$, is less similar to the global minimum energy structure. Table IX provides RMSD values comparing only the β-turn section of the Compstatin structure. In this case, the lowest energy local minima do not necessarily provide the best correspondence with the global minimum energy structure. This observation, coupled with the relatively low RMSD values between all structures, indicates that the β-turn structure is a dominant characteristic for all conformers, including the global minimum energy structure. Plots for superpositioning (backbone atoms) of the average local minimum energy structure $\overline{Compstatin}^{Local}$ and the global minimum energy structure are given in Figure 4. The superpositioning of these two structures results in characteristic RMSD values, as given in Tables VIII and IX.

4.4 Comparison with TAD: DYANA

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

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

ϕ and ψ dihedral angle values for the central β-turn residues. The problem is evidenced by the Asp^6 residue, which has ϕ-ψ values in a forbidden region of the Ramachandran plot. It appears that this may be related to clustering of the side chains in the DYANA predicted structures.

4.4.1 Including Intra-residue Restraints with DYANA

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

An additional comparison between the structural characteristics of these (DYANA) local minima and the global minimum was also performed using RMSD calculations, as given in Tables XIV and XV. These values are consistently larger than those between the average ($\overline{Compstatin}^{Local}$) and local minimum solutions structures ($< Compstatin >_i^{Local}$), and global minimum energy structure. The RMSD values not only indicate that there is significant structural difference over the entire structure (Table XIV), but also that the β-turn region (Table XV) is not a structural characteristic of the DYANA local minima. This is evidenced by the superpositioning of the lowest energy DYANA structure and the global minimum energy structure, given in Figure 5.

5 Concluding Remarks

In this work a novel and completely general method was outlined for solving the three-dimensional protein and nucleic acid structure prediction problem using conformational restraints derived from NMR data. In several ways, the method contrasts strongly with typical techniques that generally rely on the optimization of penalty-type target function using simulated annealing and molecular dynamics (plus local minimization) protocols.

One difference involves a novel reformulation of the structure prediction problem. A common characteristic of most current methods is their dependence on a penalty-type, unconstrained problem formulation, in which the objective is to minimize the sum of violation energies. In this work, the problem is formulated using nonlinear constraints, which can be individually enumerated for all or subsets of the distance restraints. In addition, the simplified potential function used by many techniques, is replaced by a full atom force field, which aids in defining the correct conformational details.

Finally, the solution technique represents another enhancement over existing methods. Rather than rely on stochastic methods for finding low energy minima, this work utilizes a deterministic method, which theoretically guarantees that the global solution will be located. This branch and bound technique, based on the αBB algorithm, has already been successfully applied to the identification of global minimum energy structures of peptides modeled by full atom force fields.

The application of this technique to the Compstatin structure prediction problem emphasizes the merits of the approach. The globally predicted structure agrees with previous results based on X-PLOR[4] when considering structural characteristics, such as the formation of a type I β-turn. However, the overall structure exhibits an improved energy, which indicates better definition of structural details. In contrast, results obtained from TAD fail to identify a type I β-turn. This is most likely attributable to the simplistic form of energy modeling and the difficulties in searching the conformational space.

Acknowledgments

The authors gratefully acknowledge financial support from the National Science Foundation, Air Force Office of Scientific Research, and the National Institutes of Health (R01 GM52032).

References

1. W. Braun and N. Go, *J. Mol. Biol.*, 186, 611–626 (1985).

2. P. Güntert, W. Braun, and K. Wüthrich, *J. Mol. Biol.*, 217, 517–530 (1991).

3. P. Güntert and K. Wüthrich, *J. Biomol. NMR*, 1, 446–456 (1991).

4. A.T. Brünger, Yale University Press, New Haven, USA, 1992.

5. P. Güntert, C. Mumenthaler, and K. Wüthrich, *J. Mol. Biol.*, 273, 283–298 (1997).

6. L. M. Rice and A. T. Brünger, *Proteins*, 19, 277–290 (1994).

7. A. T. Brünger, P. D. Adams, and L. M. Rice, *Structure*, 5, 325–336 (1997).

8. M. Nilges, *Curr. Opin. Struct. Bio.*, 6, 617–623 (1996).

9. A. E. Torda and W. F. van Gunsteren, In K. B. Lipkowitz and D. B. Boyd, editors, *Reviews in Computational Chemistry*, volume 3, page 143. VCH Publishers, 1992.

10. C. S. Adjiman, I. P. Androulakis, C. D. Maranas, and C. A. Floudas, *Comput. Chem. Eng.*, 20, S419–S424 (1996).

11. C. S. Adjiman, I. P. Androulakis, and C. A. Floudas, *Comput. Chem. Eng.*, 21, S445–S450 (1997).

12. C. S. Adjiman, I. P. Androulakis, and C. A. Floudas, *Comput. Chem. Eng.*, 22, 1159–1179 (1998).

13. C. S. Adjiman, S. Dallwig, C. A. Floudas, and A. Neumaier, *Comput. Chem. Eng.*, 22, 1137–1158 (1998).

14. I. P. Androulakis, C. D. Maranas, and C. A. Floudas, *J. Glob. Opt.*, 7, 337–363 (1995).

15. D. Morikis, N. Assa-Munt, A. Sahu, and J. D. Lambris, *Protein Sci.*, 7, 619–627 (1998).

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

17. C. A. Floudas, J. L. Klepeis, and P. M. Pardalos, In *DIMACS Series in Discrete Mathematics and Theoretical Computer science*, volume 47, pages 141–171. American Mathematical Society, 1999.

18. C.A. Floudas, In L.T. Biegler, T.F. Coleman, A.R. Conn, and F.N. Santosa, editors, *Large Scale Optimization with Applications, Part II: Optimal Design and Control*, volume 93, page 129. IMA Volumes in Mathematics and its Applications, Springer–Verlag, 1997.

19. C. D. Maranas and C. A. Floudas, *J. Chem. Phys.*, 97, 7667–7678 (1992).

20. C. D. Maranas and C. A. Floudas, *Annals of Operations Research*, 42, 85–117 (1993).

21. C. D. Maranas and C. A. Floudas, *J. Chem. Phys.*, 100, 1247–1261 (1994).

22. C. D. Maranas and C. A. Floudas, *J. Glob. Opt.*, 4, 135–170 (1994).

23. I. P. Androulakis, C. D. Maranas, and C. A. Floudas, *J. Glob. Opt.*, 11, 1–34 (1997).

24. C. D. Maranas, I. P. Androulakis, and C. A. Floudas, In *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, volume 23, page 133. American Mathematical Society, 1996.

25. J. L. Klepeis, I. P. Androulakis, M. G. Ierapetritou, and C. A. Floudas, *Comput. Chem. Eng.*, 22, 765–788 (1998).

26. J. L. Klepeis and C. A. Floudas, *J. Comp. Chem.* (1999), (in press).

27. C. S. Adjiman and C. A. Floudas, *J. Glob. Opt.*, 9, 23–40 (1996).

28. H.A. Scheraga, *PACK: Programs for Packing Polypeptide Chains*, 1996, online documentation.

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

30. J. L. Klepeis and C. A. Floudas (1999), (submitted for publication).

31. A. Sahu, B.K. Kay, and J.D. Lambris, *J. Immunol.*, 157, 884–891 (1996).

Table I: Type I β–turn for Gln^{5}–Asp^{6}–Trp^{7}–Gly^{8} segment. The first entry provides the criteria for a classic Type I β turn. $< Compstatin >$ refers to the ensemble of 21 refined structures and $\overline{Compstatin}$ refers to the average structure.

	ϕ_2 (^{o})	ψ_2 (^{o})	ϕ_3 (^{o})	ψ_3 (^{o})	C_1^{α}-C_4^{α} ($\AA$)	$=O_1$-N_4 ($\AA$)
Classic Type I β	-60 $\pm$ 30	-30 $\pm$ 30	-90 $\pm$ 30	0 $\pm$ 30	≤ 7	2-5
$< Compstatin >$	-65 $\pm$ 13	-26 $\pm$ 8	-108 $\pm$ 12	-14 $\pm$ 3	4.7 $\pm$ 0.2	3.3 $\pm$ 0.4
$\overline{Compstatin}$	-76	-23	-100	-14	4.8	3.4

Table II: Summation of distance violations D_{VIO} (column 2) and violation energy E_{VIO} (column 3) for each of the original 21 $< Compstatin >$ structures and the average structure $\overline{Compstatin}$ for backbone sequential and medium/long range NOE restraints

Structure	D_{VIO} ($\mathring{A}$)	E_{VIO} (kcal/mol)
$< Compstatin >_{1}$	5.290	129.00
$< Compstatin >_{2}$	6.686	189.77
$< Compstatin >_{3}$	5.745	145.83
$< Compstatin >_{4}$	4.749	100.21
$< Compstatin >_{5}$	4.569	114.70
$< Compstatin >_{6}$	6.545	176.63
$< Compstatin >_{7}$	5.154	129.33
$< Compstatin >_{8}$	4.269	92.14
$< Compstatin >_{9}$	5.708	150.63
$< Compstatin >_{10}$	5.492	152.89
$< Compstatin >_{11}$	5.565	163.72
$< Compstatin >_{12}$	5.204	129.98
$< Compstatin >_{13}$	6.000	169.76
$< Compstatin >_{14}$	5.679	164.39
$< Compstatin >_{15}$	5.036	107.97
$< Compstatin >_{16}$	5.298	137.34
$< Compstatin >_{17}$	6.848	211.47
$< Compstatin >_{18}$	6.349	206.90
$< Compstatin >_{19}$	4.278	113.85
$< Compstatin >_{20}$	5.160	114.31
$< Compstatin >_{21}$	6.589	173.11
$\overline{Compstatin}$	6.919	205.90

Table III: Local minimization results for the ECEPP/3 $< Compstatin >$ and $\overline{Compstatin}$ starting structures

Local minimum	D$_{VIO}$ ($\AA$)	E$_{ECEPP/3}$ (kcal/mol)
$< Compstatin >_1^{Local}$	6.547	-37.230
$< Compstatin >_2^{Local}$	6.963	-71.613
$< Compstatin >_3^{Local}$	6.293	-17.120
$< Compstatin >_4^{Local}$	6.727	-17.927
$< Compstatin >_5^{Local}$	7.343	-41.558
$< Compstatin >_6^{Local}$	6.622	-58.095
$< Compstatin >_7^{Local}$	6.481	-54.068
$< Compstatin >_8^{Local}$	7.064	-36.832
$< Compstatin >_9^{Local}$	7.120	-67.653
$< Compstatin >_{10}^{Local}$	5.873	-41.685
$< Compstatin >_{11}^{Local}$	7.185	-61.843
$< Compstatin >_{12}^{Local}$	7.056	-42.540
$< Compstatin >_{13}^{Local}$	6.510	-43.081
$< Compstatin >_{14}^{Local}$	6.847	-47.396
$< Compstatin >_{15}^{Local}$	6.789	-35.095
$< Compstatin >_{16}^{Local}$	6.035	-41.594
$< Compstatin >_{17}^{Local}$	6.540	-62.537
$< Compstatin >_{18}^{Local}$	6.764	-54.813
$< Compstatin >_{19}^{Local}$	7.158	-65.825
$< Compstatin >_{20}^{Local}$	7.348	-35.491
$< Compstatin >_{21}^{Local}$	6.832	-68.704
$\overline{Compstatin}^{Local}$	6.392	-52.283

Table IV: RMSD values for full Compstatin structures using all heavy atoms. Column 2 compares the original PDB structure ($< Compstatin >_i$) to the average Compstatin PDB structure ($\overline{Compstatin}$). Column 3 compares the original PDB structure ($< Compstatin >_i$) to the ECEPP/3 local minimum using this structure as a starting point ($< Compstatin >_i^{Local}$). Column 4 compares this local minimum structure ($< Compstatin >_i^{Local}$) to the local minimum for the average Compstatin structure ($\overline{Compstatin}^{Local}$).

Structure	$\overline{Compstatin}$-Original	Original-Local	Local-$\overline{Compstatin}$
$< Compstatin >_1$	2.372	1.988	2.844
$< Compstatin >_2$	1.979	3.671	2.021
$< Compstatin >_3$	2.445	3.415	2.865
$< Compstatin >_4$	1.910	2.235	2.249
$< Compstatin >_5$	2.185	3.162	2.386
$< Compstatin >_6$	2.438	3.317	2.513
$< Compstatin >_7$	1.934	5.074	4.472
$< Compstatin >_8$	2.268	3.058	3.047
$< Compstatin >_9$	2.030	3.756	2.433
$< Compstatin >_{10}$	2.387	4.176	3.363
$< Compstatin >_{11}$	2.567	3.275	2.662
$< Compstatin >_{12}$	2.314	3.737	2.509
$< Compstatin >_{13}$	2.000	3.092	2.083
$< Compstatin >_{14}$	2.148	3.314	2.915
$< Compstatin >_{15}$	1.847	2.332	2.024
$< Compstatin >_{16}$	2.089	2.421	2.349
$< Compstatin >_{17}$	2.438	3.849	3.073
$< Compstatin >_{18}$	2.480	3.422	2.211
$< Compstatin >_{19}$	2.142	4.104	2.561
$< Compstatin >_{20}$	2.145	2.315	2.547
$< Compstatin >_{21}$	2.305	3.596	2.257
$\overline{Compstatin}$	—	2.773	—

Table V: RMSD values for full Compstatin structures using only backbone atoms (N,C^{α},C'). Column 2 compares the original PDB structure ($< Compstatin >_i$) to the average Compstatin PDB structure ($\overline{Compstatin}$). Column 3 compares the original PDB structure ($< Compstatin >_i$) to the ECEPP/3 local minimum using this structure as a starting point ($< Compstatin >_i^{Local}$). Column 4 compares this local minimum structure ($< Compstatin >_i^{Local}$) to the local minimum for the average Compstatin structure ($\overline{Compstatin}^{Local}$).

Structure	$\overline{Compstatin}$-Original	Original-Local	Local-$\overline{Compstatin}$
$< Compstatin >_1$	1.510	1.442	2.301
$< Compstatin >_2$	1.229	2.976	1.681
$< Compstatin >_3$	1.740	2.820	2.496
$< Compstatin >_4$	0.978	1.625	1.192
$< Compstatin >_5$	1.379	2.361	1.008
$< Compstatin >_6$	1.354	2.832	2.297
$< Compstatin >_7$	1.144	3.938	3.331
$< Compstatin >_8$	1.662	1.565	1.860
$< Compstatin >_9$	1.437	2.782	1.394
$< Compstatin >_{10}$	1.968	3.052	2.500
$< Compstatin >_{11}$	1.427	2.366	1.746
$< Compstatin >_{12}$	1.542	2.757	1.576
$< Compstatin >_{13}$	1.539	2.072	0.898
$< Compstatin >_{14}$	1.561	2.833	1.728
$< Compstatin >_{15}$	1.151	1.569	1.249
$< Compstatin >_{16}$	1.420	2.058	1.440
$< Compstatin >_{17}$	1.458	3.176	1.732
$< Compstatin >_{18}$	1.970	2.366	1.071
$< Compstatin >_{19}$	1.662	2.898	1.553
$< Compstatin >_{20}$	1.560	1.901	1.465
$< Compstatin >_{21}$	1.911	2.608	1.165
$\overline{Compstatin}$	—	1.633	—

Table VI: RMSD values for the β-turn regions (Residues 5 through 8) using all heavy atoms. Column 2 compares the original PDB structure ($< Compstatin >_i$) to the average Compstatin PDB structure ($\overline{Compstatin}$). Column 3 compares the original PDB structure ($< Compstatin >_i$) to the ECEPP/3 local minimum using this structure as a starting point ($< Compstatin >_i^{Local}$). Column 4 compares this local minimum structure ($< Compstatin >_i^{Local}$) to the local minimum for the average Compstatin structure ($\overline{Compstatin}^{Local}$).

Structure	$\overline{Compstatin}$-Original	Original-Local	Local-$\overline{Compstatin}$
$< Compstatin >_1$	0.995	0.531	0.865
$< Compstatin >_2$	0.765	0.718	0.617
$< Compstatin >_3$	0.876	0.556	0.990
$< Compstatin >_4$	0.759	0.715	0.827
$< Compstatin >_5$	0.711	0.520	0.800
$< Compstatin >_6$	0.813	0.918	1.039
$< Compstatin >_7$	0.437	1.693	1.755
$< Compstatin >_8$	1.025	0.626	0.825
$< Compstatin >_9$	0.633	0.547	0.394
$< Compstatin >_{10}$	0.521	0.579	0.922
$< Compstatin >_{11}$	0.835	0.625	0.590
$< Compstatin >_{12}$	0.941	0.638	0.933
$< Compstatin >_{13}$	0.728	0.767	0.542
$< Compstatin >_{14}$	0.814	0.780	0.813
$< Compstatin >_{15}$	0.713	0.630	0.569
$< Compstatin >_{16}$	0.818	1.003	0.515
$< Compstatin >_{17}$	0.830	0.595	0.901
$< Compstatin >_{18}$	0.667	0.704	0.564
$< Compstatin >_{19}$	0.786	0.492	0.657
$< Compstatin >_{20}$	0.668	0.771	0.662
$< Compstatin >_{21}$	1.062	1.318	0.660
$\overline{Compstatin}$	—	0.758	—

Table VII: RMSD values for the β-turn regions (Residues 5 through 8) using only backbone atoms (N,C^{α},C'). Column 2 compares the original PDB structure ($< Compstatin >_i$) to the average Compstatin PDB structure ($\overline{Compstatin}$). Columns 3 compares the original PDB structure ($< Compstatin >_i$) to the ECEPP/3 local minimum using this structure as a starting point ($< Compstatin >_i^{Local}$). Column 4 compares this local minimum structure ($< Compstatin >_i^{Local}$) to the local minimum for the average Compstatin structure ($\overline{Compstatin}^{Local}$).

Structure	$\overline{Compstatin}$-Original	Original-Local	Local-$\overline{Compstatin}$
$< Compstatin >_1$	0.189	0.235	0.170
$< Compstatin >_2$	0.231	0.235	0.106
$< Compstatin >_3$	0.106	0.239	0.183
$< Compstatin >_4$	0.266	0.313	0.080
$< Compstatin >_5$	0.180	0.172	0.159
$< Compstatin >_6$	0.355	1.322	0.375
$< Compstatin >_7$	0.210	1.026	1.059
$< Compstatin >_8$	0.194	0.142	0.092
$< Compstatin >_9$	0.062	0.222	0.094
$< Compstatin >_{10}$	0.254	1.221	0.335
$< Compstatin >_{11}$	0.433	0.368	0.074
$< Compstatin >_{12}$	0.215	0.376	0.388
$< Compstatin >_{13}$	0.291	0.306	0.079
$< Compstatin >_{14}$	0.263	0.418	0.515
$< Compstatin >_{15}$	0.167	0.266	0.116
$< Compstatin >_{16}$	0.115	0.414	0.384
$< Compstatin >_{17}$	0.173	0.199	0.199
$< Compstatin >_{18}$	0.193	0.158	0.067
$< Compstatin >_{19}$	0.185	0.248	0.143
$< Compstatin >_{20}$	0.202	0.361	0.318
$< Compstatin >_{21}$	0.336	0.480	0.172
$\overline{Compstatin}$	——	0.197	——

Table VIII: RMSD values for full Compstatin structures. Column 2 reports RMSD using all heavy atoms, while 3 accounts for only backbone atoms (N,C^{α},C'). Both columns compare the ECEPP/3 local minimum structures ($< Compstatin >_i^{Local}$ and $\overline{Compstatin}^{Local}$) to the global minimum Compstatin PDB structure ($Compstatin^{Global}$).

Structure	Heavy Atoms	Backbone Atoms
$< Compstatin >_1$	4.106	3.352
$< Compstatin >_2$	2.205	1.220
$< Compstatin >_3$	2.742	2.265
$< Compstatin >_4$	2.579	1.988
$< Compstatin >_5$	2.925	1.541
$< Compstatin >_6$	2.513	2.080
$< Compstatin >_7$	4.866	3.314
$< Compstatin >_8$	2.906	2.584
$< Compstatin >_9$	1.287	0.953
$< Compstatin >_{10}$	2.609	2.317
$< Compstatin >_{11}$	1.365	1.156
$< Compstatin >_{12}$	1.824	1.376
$< Compstatin >_{13}$	2.497	1.638
$< Compstatin >_{14}$	2.676	2.110
$< Compstatin >_{15}$	3.475	2.359
$< Compstatin >_{16}$	3.089	2.239
$< Compstatin >_{17}$	1.385	1.074
$< Compstatin >_{18}$	1.898	1.651
$< Compstatin >_{19}$	1.304	1.046
$< Compstatin >_{20}$	3.593	2.346
$< Compstatin >_{21}$	1.565	1.086
$\overline{Compstatin}$	2.778	1.625

Table IX: RMSD values for the β-turn regions (Residues 5 through 8). Column 2 reports RMSD using all heavy atoms, while 3 accounts for only backbone atoms (N,C$^{\alpha}$,C'). Both columns compare the ECEPP/3 local minimum structures ($< Compstatin >_i^{Local}$ and $\overline{Compstatin}^{Local}$) to the global minimum Compstatin PDB structure ($Compstatin^{Global}$).

Structure	Heavy Atoms	Backbone Atoms
$< Compstatin >_1$	1.061	0.288
$< Compstatin >_2$	0.510	0.271
$< Compstatin >_3$	1.114	0.244
$< Compstatin >_4$	1.214	0.259
$< Compstatin >_5$	0.771	0.317
$< Compstatin >_6$	1.160	0.358
$< Compstatin >_7$	1.766	0.854
$< Compstatin >_8$	1.267	1.185
$< Compstatin >_9$	0.792	0.271
$< Compstatin >_{10}$	0.952	0.268
$< Compstatin >_{11}$	0.579	0.325
$< Compstatin >_{12}$	1.243	0.391
$< Compstatin >_{13}$	0.535	0.284
$< Compstatin >_{14}$	1.147	0.526
$< Compstatin >_{15}$	0.565	0.298
$< Compstatin >_{16}$	0.974	0.211
$< Compstatin >_{17}$	0.918	0.284
$< Compstatin >_{18}$	0.607	0.295
$< Compstatin >_{19}$	0.543	0.288
$< Compstatin >_{20}$	0.763	0.194
$< Compstatin >_{21}$	0.528	0.306
$\overline{Compstatin}$	0.774	0.295

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

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

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

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

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

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

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

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

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

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

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

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

Figure legends

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

Figure 2: Superposition of $\overline{Compstatin}_{Orig}$ structure (in light grey) and corresponding ECEPP/3 structure (in black) using calculated dihedral angles ($\overline{Compstatin}_{ECEPP}$).

Figure 3: Superposition of $< Compstatin >^{Local}_{13}$ (in black) and $\overline{Compstatin}^{Local}$ (in light grey) structures. The left panel shows the full (all atom) structure, while the right panel compares only the β-turn region.

Figure 4: 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.

Figure 5: Superposition of global minimum (in black) and $Compstatin^{DYANA}_{1c}$ (in grey) structures. The left panel shows the full (backbone atom) structure, while the right panel compares only the β-turn region.

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

Figure 1:

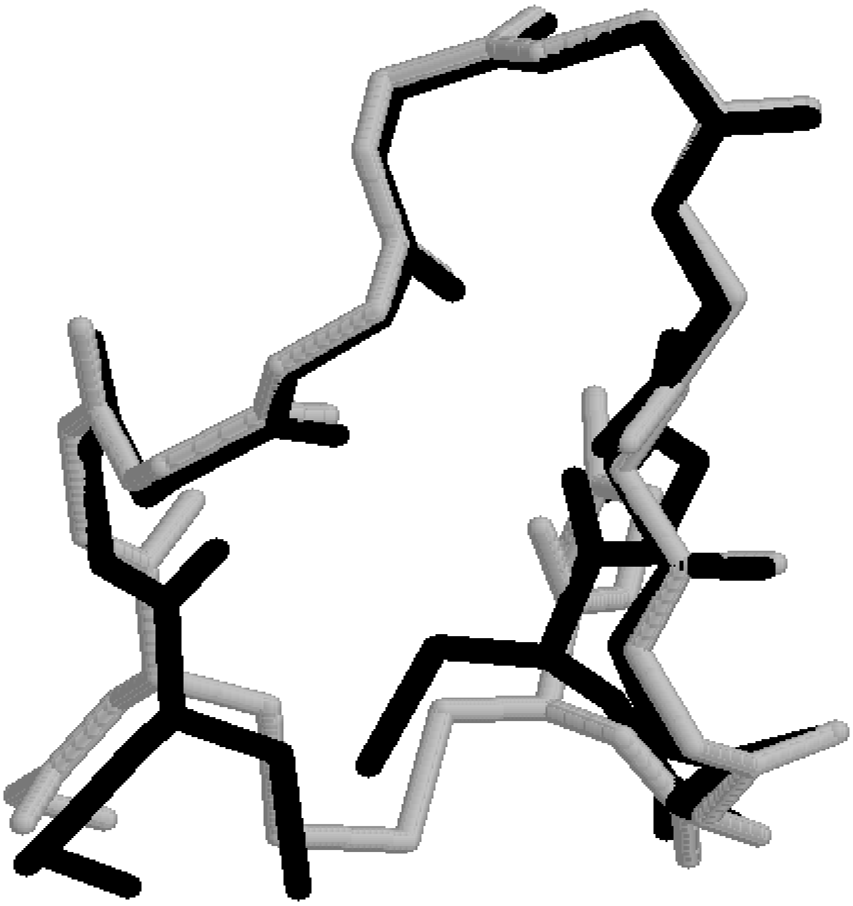

Figure 2:

Figure 3:

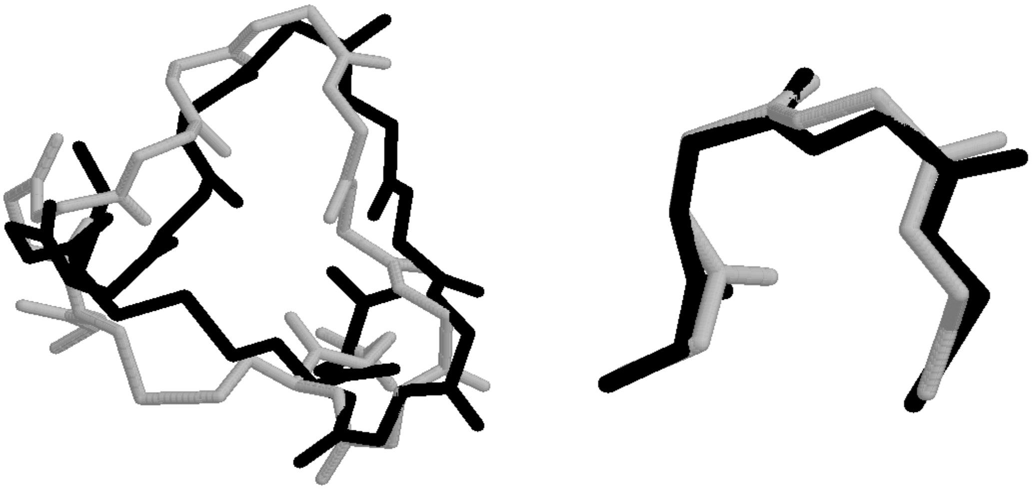

Figure 4:

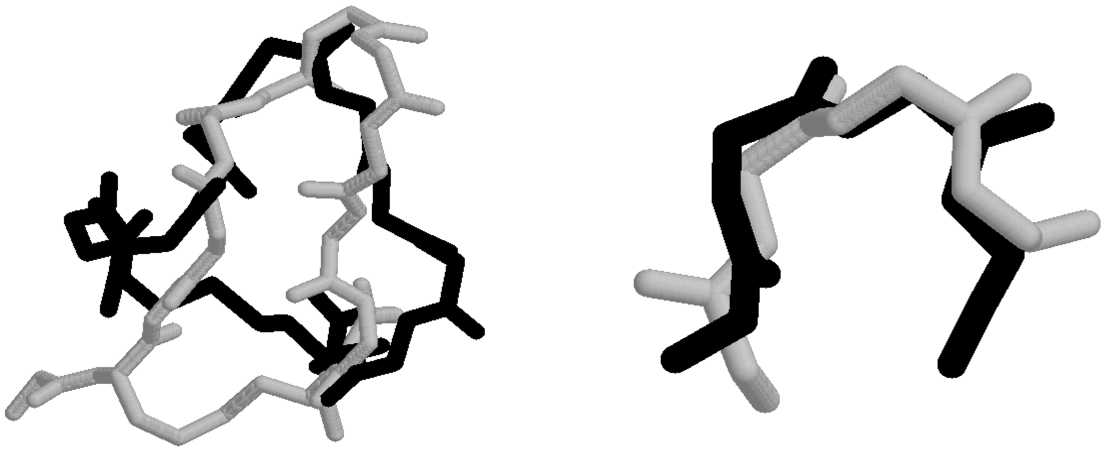

Figure 5: