

# Prediction of Oligopeptide Conformations via Deterministic Global Optimization

I.P. ANDROULAKIS

*Corporate Research Science Laboratories, Exxon Research and Engineering Company, Route 22 East, Annandale, NJ 08801.*

C.D. MARANAS

*Department of Chemical Engineering, The Pennsylvania State University, 158 Fenske Laboratory, University Park, PA 16802.*

C.A. FLOUDAS\*

*Department of Chemical Engineering, Princeton University, Princeton, N.J. 08544 5269*

**Abstract.** A deterministic global optimization method is described for identifying the global minimum potential energy conformation of oligopeptides. The ECEPP/3 detailed potential energy model is utilized for describing the energetics of the atomic interactions posed in the space of the peptide dihedral angles. Based on previous work on the microcluster and molecular structure determination [21, 22, 23, 24], a procedure for deriving convex lower bounding functions for the total potential energy function is developed. A procedure that allows the exclusion of domains of the  $(\phi, \psi)$  space based on the analysis of experimentally determined native protein structures is presented. The reduced disjoint sub-domains are appropriately combined thus defining the starting regions for the search. The proposed approach provides valuable information on (i) the global minimum potential energy conformation, (ii) upper and lower bounds of the global minimum energy structure and (iii) low energy conformers close to the global minimum one. The proposed approach is illustrated with Ac-Ala<sub>4</sub>-Pro-NHMe, Met-enkephalin, Leu-enkephalin, and Decaglycine.

**Keywords:** Protein folding, deterministic global optimization.

## 1. Introduction

The *protein folding* problem is one of the most important problems in biochemistry. Predicting how a protein would fold is of paramount academic and industrial interest. Many products of the biotechnology industry are novel proteins. Knowledge of how the protein would fold would allow one to predict and fine-tune its chemical and biological properties. This would greatly simplify the tasks of interpreting data collected by the human genome project, understanding the mechanisms of hereditary and infectious diseases, designing drugs with specific therapeutical properties, and growing biological polymers with specific material properties. Although, small molecules exist in an ensemble of low-energy conformations [41], proteins in their biologically active (native) state exist in a well-defined, recognizable conformation with small fluctuations around this average. There is considerable evidence that

---

\*AUTHOR TO WHOM ALL CORRESPONDENCE SHOULD BE ADDRESSED.

proteins do fold spontaneously, both *in vivo* and *in vitro*, into their native conformations. This native conformation is uniquely determined by the amino acid sequence, environment (i.e., solvent) and conditions (i.e., temperature, pH). According to the thermodynamic hypothesis [8] this most stable protein conformation corresponds to the one with the lowest (global) minimum free energy. This implies that at a finite temperature the probabilities of occurrence of conformation states will be significantly different than zero only for a distinctly unique low energy low potential energy conformation. This work addresses this problem at various levels. First of all, a deterministic global optimization approach that identifies an  $\epsilon$ -global minimum potential energy conformation will be proposed. Furthermore, rigorous upper and lower bounds on the global minimum total potential energy will be provided. Finally, a number of low energy conformers will be identified.

A *protein* is a polymer chain composed by a sequence of various amino acid residues connected with peptide bonds. Proteins in nature are composed of only twenty different amino acid residues. Instead of specifying the coordinate vector for all atoms in a protein, one can specify all bond lengths, covalent bond angles and dihedral angles. Under biological conditions, the bond lengths and bond angles are fairly rigid and thus can be assumed to be fixed at their equilibrium values. Under this assumption, the dihedral angles along the backbone fully determine the geometric shape of the folded protein. The names of the dihedral angles of a folded protein chain follow a standard nomenclature. The dihedral angle between the normals of the planes formed by atoms  $C'_{i-1}N_iC_{\alpha,i}$  and  $N_iC_{\alpha,i}C'_i$  respectively is denoted as  $\phi_i$  where  $i-1$  and  $i$  are two adjacent amino acid residues. The one defined by planes  $R_iC_{\alpha,i}C'_i$  and  $C_{\alpha,i}C'_iN_{i+1}$  respectively is denoted as  $\psi_i$  where  $i$  and  $i+1$  are two adjacent amino acid residues. Also  $\omega_i$  is the dihedral angle defined by the planes  $C_{\alpha,i}C'_iN_{i+1}$  and  $C'_iN_{i+1}C_{\alpha,i+1}$ . The letter  $\chi$  is utilized to denote the dihedral angles which are associated with the side groups  $R_i$ . Also the letter  $\theta$  is used to name the dihedral angles associated with the two end groups. Figure 1 pictorially illustrates these conventions. Excellent surveys of the key issues and approaches for predicting structures of oligopeptides, polypeptides, and proteins are reported in [41], [35], as well as in [37].

Polypeptide folding calculations typically employ an empirically derived set of potential energy contributions for approximating the true potential function of the protein system. This set of potential energy contributions, called the *force field*, contains adjustable parameters that are selected in a such a way as to provide the best possible agreement with experimental data. The main assumption introduced in molecular mechanics is that every parameter is associated with a *specific interaction* rather than a specific molecule (*transferability assumption*). These parameters are bond lengths; covalent bond angles; bond stretching, bending, or rotating constants; non-bonded atom interaction constants, etc. Thus, whenever a specific interaction is present, the same value for the parameter can be used even if this interaction occurs in different molecules [16]. Many different parameterizations have been proposed for approximating the *force field* in protein folding calculations. Some of the most popular ones are: ECEPP [27, 28, 29], MM2 [3], ECEPP/2

[33], CHARMM [10], DISCOVER [12], AMBER [47, 48], GROMOS87 [44], ENCAD [19], MM3 [4], and ECEPP/3 [34]. In this work the ECEPP/3 [34] detailed potential model is utilized. In this potential model, it is assumed that the covalent bond lengths and angles are fixed at their equilibrium values and the conformational energy is treated as the sum of electrostatic, nonbonded, hydrogen bond and torsional contributions, a loop closing potential if the polypeptide contains one or more intramolecular disulfide bonds, plus the fixed internal conformational energy of the pyrrolidine ring for each propyl or hydroxypropyl residue contained in the peptide chain. The latter is implemented by allowing the user two choices for the pyrrolidine geometry: *Up* or *Down*. In short, the potential function that ECEPP/3 generates includes the following terms :

$$\begin{aligned}
U &= \sum_{(i,j) \in \mathcal{E}\mathcal{S}} 332.0 \frac{q_i q_j}{D r_{ij}} \quad (\text{Electrostatic}) \\
&+ \sum_{(i,j) \in \mathcal{N}\mathcal{B}} F \frac{A}{r_{ij}^{12}} - \frac{C}{r_{ij}^6} \quad (\text{Nonbonded}) \\
&+ \sum_{(hx) \in \mathcal{H}\mathcal{A}} F \frac{A'}{r_{hx}^{12}} - \frac{B}{r_{hx}^{10}} \quad (\text{Hydrogen bonding}) \\
&+ \sum_{k \in \mathcal{T}\mathcal{O}\mathcal{R}} \left(\frac{U_o}{2}\right) (1 \pm \cos n_k \theta_k) \quad (\text{Torsional}) \\
&+ \sum_{i \in \mathcal{L}\mathcal{O}\mathcal{O}\mathcal{P}} B_L \sum_{i_i=1}^{i_i=3} (r_{i_i} - r_{i_o})^2 \quad (\text{Cystine Loop-Closing}) \\
&+ \sum_{i \in \mathcal{L}\mathcal{O}\mathcal{O}\mathcal{P}} A_L (r_{a_i} - r_{a_o})^2 \quad (\text{Cystine Torsional})
\end{aligned}$$

All constants have been appropriately estimated through fitting of experimental data, and are reported in ECEPP/3 [34].

## 2. Problem Formulation

The potential energy minimization problem can be formulated as a nonconvex nonlinear optimization problem. Let  $i = 1, \dots, N_{RES}$  be an indexed set describing the sequence of amino acid residues in the peptide chain. This implies that there are  $\phi_i, \psi_i, \omega_i$ ,  $i = 1, \dots, N_{RES}$  dihedral angles along the backbone of the peptide chain. Also let  $k = 1, \dots, K^i$  be an index set denoting the dihedral angles of the side group of the  $i^{th}$  residue and  $j = 1, \dots, J^N$  be an index set denoting the dihedral angles of the amino end group and  $j = 1, \dots, J^C$  be an index set denoting the dihedral angles of the carboxyl end group respectively. Over these index sets one can define the side group dihedral angles  $\chi_i^k$ ,  $i = 1, \dots, N_{RES}$ ,  $k = 1, \dots, K^i$ , the amino  $\theta_j^N$ ,  $j = 1, \dots, J^N$  and carboxyl  $\theta_j^C$ ,  $j = 1, \dots, J^C$  end group dihedral

angles respectively. Based on these definitions the potential model minimization energy problem can be formulated as follows:

$$\begin{aligned}
 \min \quad & U(\phi_i, \psi_i, \omega_i, \chi_i^k, \theta_j^N, \theta_j^C) \\
 \text{subject to} \quad & -\pi \leq \phi_i \leq \pi, \quad i = 1, \dots, N_{RES} \\
 & -\pi \leq \psi_i \leq \pi, \quad i = 1, \dots, N_{RES} \\
 & -\pi \leq \omega_i \leq \pi, \quad i = 1, \dots, N_{RES} \\
 & -\pi \leq \chi_i^k \leq \pi, \quad i = 1, \dots, N_{RES}, \quad k = 1, \dots, K^i \\
 & -\pi \leq \theta_j^N \leq \pi, \quad j = 1, \dots, J_N \\
 & \pi \leq \theta_j^C \leq \pi, \quad j = 1, \dots, J_C
 \end{aligned} \tag{1}$$

Here  $U$  is the expression for the total potential energy as a function of the peptide dihedral angles. The specific expressions comprising  $U$  have been described in the previous section.

Note that  $U$  is a nonconvex function of these dihedral angles involving a large [46] number of local minima, even for small peptide systems. These local minima correspond to metastable states of the polypeptide chain. A single global minimum defines the energetically most favorable peptide conformation. A large variety of procedures have been developed for searching the multidimensional peptide conformational space in an attempt to focus in the neighborhood containing the global minimum. These procedures draw from one or more of the following basic ideas: (i) decomposition of the conformation calculations, (ii) use of statistical and/or heuristic conformational information, (iii) further simplifications on potential model, (iv) stochastic search procedures, (v) mathematical transformations. Most methods attempt to locate this point by tracing, deterministically or stochastically, single or multiple paths on the potential energy surface conjecturing that some of them will converge to the global minimum potential energy point. The key limitation of these methods is that the obtained conformations depend heavily on the supplied initial conformation expressing the bias of the researcher towards which is the most appropriate conformation. This is why, in practice, many trial geometries need to serve as initial points in an attempt to lessen the initial point dependence. However, there is no guarantee that important conformations are not overlooked. The need for a method that can guarantee convergence to the global minimum potential energy conformation motivated our initial effort to introduce such a method for microclusters [21, 22], and small acyclic molecules [24, 23] allowing for nonbonded atomic pair interactions. The approach was subsequently extended to include realistic potential models like ECEPP/3, [25]. It was shown that this approach could efficiently identify apart from the global minimum configuration, low energy conformers as well as upper and lower bounds of the global minimum potential energy for systems composed of single residues as well as di-peptides.

The rest of the paper is structured as follows. Sections 3.1 and 3.2 provide a brief introduction to the deterministic global optimization algorithm,  $\alpha$ BB. Subsequently, in Section 3.3 the problem of identifying the global minimum total potential energy is formulated and addressed within the framework of  $\alpha$ BB. We proceed by discussing in Sections 4.1 and 4.2 a domain partitioning strategy for the identification of important domains within which the global minimum is most likely to be found. A novel approach is proposed in Section 4.3 for incorporating such a partitioning within the  $\alpha$ BB framework and computational studies on four oligopeptides, namely, *(Ala)<sub>4</sub> Pro*, *Met-enkephalin*, *Leu-enkephalin*, and *Decuglycine* are presented in Section 5.

### 3. Global Minimization of Potential Energy

#### 3.1. Deterministic Global Optimization, $\alpha$ BB

In this section we will present a brief overview of a branch-and-bound global optimization method based on the concept of the *difference of convex functions*, denoted as  $\alpha$ BB.

The general optimization problem addressed can be formulated as the following constrained nonlinear optimization problem involving only continuous variables.

$$\begin{aligned}
 & \min_{\mathbf{x}} && f(\mathbf{x}) && \text{(P0)} \\
 \text{subject to} & && h_j(\mathbf{x}) = 0, && j = 1, \dots, M \\
 & && g_k(\mathbf{x}) \leq 0, && k = 1, \dots, K \\
 & && \mathbf{x}^L \leq \mathbf{x} \leq \mathbf{x}^U \\
 & && f, \mathbf{g}, \mathbf{h} \in \mathcal{C}^2
 \end{aligned}$$

Here  $\mathbf{x}$  denotes the vector of variables,  $f(\mathbf{x})$  is the nonlinear objective function,  $h_j(\mathbf{x})$  is the set of nonlinear equality constraints, and  $g_k(\mathbf{x})$ ,  $k = 1, \dots, K$  is the set of nonlinear inequality constraints. Formulation (P0) in general corresponds to a nonconvex optimization problem possibly involving multiple local and disconnected feasible regions. Existing path following techniques cannot consistently locate the global minimum solution of (P0) even if a multi-start procedure are utilized. For special cases of (P0) efficient algorithms have been proposed for locating the global minimum solution. For the general case, however, of minimizing a nonconvex function subject to a set of nonconvex equality and inequality constraints there has been comparatively little work in deriving global optimization methods and tools.

The  $\alpha$ BB global optimization approach is based on the convex relaxation of the original nonconvex formulation (P0). This requires the convex lower bounding of all nonconvex expressions appearing in (P0). These terms can be partitioned into

three classes: (i) convex, (ii) nonconvex of special structure, and (iii) nonconvex of generic structure.

Clearly, no convex lower bounding action is required for convex functions. For nonconvex terms of special structure (e.g., bilinear, univariate concave functions), tight specialized convex lower bounding schemes already exist and therefore can be utilized. The general case of constrained nonconvex nonlinear optimization problems as stated in (P0) is studied in [5]. Since the potential energy minimization of oligopeptides problem has the mathematical structure of (1) (i.e., nonlinear objective function subject to box constraints), we will retain only those parts of the formulation that pertain to the problem addressed in this paper. The terms appearing in the objective function are rewritten equivalently as follows:

$$\begin{aligned} \min_{\mathbf{x}} \quad & C^0(\mathbf{x}) + \sum_{k \in \mathcal{K}^0} NC_k^0(\mathbf{x}) & (\mathbf{P}) \\ \text{subject to} \quad & \mathbf{x}^L \leq \mathbf{x} \leq \mathbf{x}^U \end{aligned}$$

where  $NC_k^0(\mathbf{x})$  with  $\mathbf{x} \in \{x_i : i \in \mathcal{N}_k^0\}$  and  $C^0(\mathbf{x})$  denotes a part of the objective function that may be identified as convex.

A convex relaxation of (P) can be constructed by replacing each generic nonconvex term,  $NC_k^j(\mathbf{x})$ , with one or more convex lower bounding functions. The convex lower bounding of the generic nonconvex terms  $NC_k^j$  is motivated by the approach introduced in [23] where it was shown that by considering the dual formulation of a difference of convex functions problem results in a convex lower bounding function  $NC_k^{0,conv}$  equivalent to augmenting the original nonconvex expression with the addition of a separable convex quadratic function of  $(x_i, i \in \mathcal{N}_k^0)$ .

$$\begin{aligned} & NC_k^{0,conv}(\mathbf{x}) - NC_k^0(\mathbf{x}) \\ & + \sum_{i \in \mathcal{N}_k^j} \alpha_{i,k}^0(\mathbf{x}^L, \mathbf{x}^U) (x_i^L - x_i) (x_i^U - x_i), \quad k \in \mathcal{K}^j \end{aligned}$$

where  $\alpha_{i,k}^0(\mathbf{x}^L, \mathbf{x}^U) \geq \max \left\{ 0, -\frac{1}{2} \min_{\mathbf{x}^L \leq \mathbf{x} \leq \mathbf{x}^U} \lambda(\mathbf{x}) \right\}$

Note that  $\alpha_{i,k}^0$  are nonnegative parameters which must be greater or equal to the negative one half of the minimum eigenvalue of the Hessian matrix of  $NC_k^{0,conv}$  over  $x_i^L \leq x_i \leq x_i^U$ ,  $i \in \mathcal{N}_k^0$ . These parameters  $\alpha_{i,k}^0$  can be estimated either through the solution of an optimization problem or by using the concept of the measure of a matrix [23]. The effect of adding the extra separable quadratic term on the generic nonconvex terms is to construct new convex functions by overpowering the nonconvexity characteristics of the original nonconvex terms with the addition of the terms  $2\alpha_{i,k}^0$  to all of their eigenvalues. The new function  $NC_k^{0,conv}$  defined over the rectangular domains  $x_i^L \leq x_i \leq x_i^U$ ,  $i \in \mathcal{N}_k^0$  involves a number of important properties. These properties are as follows:

- Property 1:  $NC_k^{0,conv}$  is a valid *underestimator* of  $NC_k^0$ .

- Property 2:  $NC_k^{0,conv}(\mathbf{x})$  matches  $NC_k^0(\mathbf{x})$  at all corner points.
- Property 3:  $NC_k^{0,conv}(\mathbf{x})$  is *convex* in  $x_i \in [x_i^L, x_i^U]$ ,  $i \in \mathcal{N}_k^0$ .
- Property 4: The maximum separation between the nonconvex term of generic structure  $NC_k^{0,conv}$  and its convex relaxation  $NC_k^0$  is *bounded* and proportional to the positive parameters  $\alpha_{i,k}^0$  and to the square of the diagonal of the current box constraints.
- Property 5: 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.

Clearly, the smaller the values of the positive parameters  $\alpha_{i,k}^0$ , the narrower the separation between the original nonconvex terms and their respective convex relaxations will be. Therefore fewer iterations will also be required for convergence. To this end, customized  $\alpha$  parameters can be defined for each variable, and term. Furthermore, an updating procedure for the  $\alpha$ 's as the size of the partition elements decreases allows for substantial improvement in the convergence process.

Based on the aforementioned convex lower bounding procedures a convex relaxation **(R)** of **(P)** is proposed.

$$\begin{aligned}
 \min_{\mathbf{x}} C^0(\mathbf{x}) + \sum_{k \in \mathcal{K}^0} NC_k^0(\mathbf{x}) & \quad \mathbf{(R)} \\
 + \sum_{i \in \mathcal{N}_k^0} \alpha_{i,k}^0(\mathbf{x}^L, \mathbf{x}^U) (x_i^L - x_i) (x_i^U - x_i) & \\
 \mathbf{x}^L \leq \mathbf{x} \leq \mathbf{x}^U & \\
 \text{and } NC_k^0(\mathbf{x}) \text{ with } \mathbf{x} \in \{x_i : i \in \mathcal{N}_k^0\}, &
 \end{aligned}$$

Formulation **(R)** is a convex programming problem whose global minimum solution can be routinely found with existing local optimization solvers such as MINOS5.4 [26]. Formulation **(R)** is a relaxation of **(P)** and therefore its solution is a valid lower bound on the global minimum solution of **(P)**.

In the next section, we will see how this convex lower bounding formulation **(R)** can be utilized in a branch and bound framework for locating the global minimum solution of **(P)**.

### 3.2. Global Optimization Algorithm, $\alpha$ BB

A global optimization procedure,  $\alpha$ BB, is proposed for locating the global minimum solution of **(P)** based on the refinement of converging lower and upper bounds. Lower bounds are obtained through the solution of convex programming problems **(R)** and upper bounds based on the solution of **(P)** with local methods.

As it has been discussed in the previous section, the maximum separation between the the generic nonconvex terms and their respective convex lower bounding

functions is proportional to the square of the diagonal of the rectangular partition element. Furthermore, as the size of the rectangular domains approaches zero, these maximum separations go to zero as well. This implies that as the current box constraints  $[\mathbf{x}^L, \mathbf{x}^U]$  collapse into a point; (i) the maximum separation between the original objective function of  $(\mathbf{P})$  and its convex relaxation in  $(\mathbf{R})$  becomes zero; and (ii) by the same argument, the maximum separation between the original constraint set in  $(\mathbf{P})$  and the one in  $(\mathbf{R})$  goes to zero as well. This implies that for every positive number  $\epsilon_f$  and  $\mathbf{x}$  there always exists a positive number  $\delta$  such that by reducing the rectangular region  $[\mathbf{x}^L, \mathbf{x}^U]$  around  $\mathbf{x}$  so as  $\|\mathbf{x}^U - \mathbf{x}\| \leq \delta$  differences between the feasible region of the original problem  $(\mathbf{P})$  and its convex relaxation  $(\mathbf{R})$  become less than  $\epsilon_f$ . Therefore, any feasible point  $\mathbf{x}^c$  of problem  $(\mathbf{R})$  (even the global minimum solution) becomes at least  $\epsilon_f$ -feasible for problem  $(\mathbf{P})$  by sufficiently tightening the bounds on  $\mathbf{x}$  around this point.

The next step, after establishing an upper and a lower bound on the global minimum, is to refine them. This is accomplished by successively partitioning the initial rectangular region into smaller ones. 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 in formulation  $(\mathbf{P})$ . The partitioning strategy involves the successive subdivision of a rectangle into two sub-rectangles by halving on the middle point of the longest side of the initial rectangle (bisection). Therefore, at each iteration a lower bound of the objective function of  $(\mathbf{P})$  is simply the minimum over all the minima of problem  $(\mathbf{R})$  in every sub-rectangle composing the initial rectangle. Therefore, a straightforward (bound improving) way of tightening the lower bound is to halve at each iteration, only the sub-rectangle responsible for the infimum of the minima of  $(\mathbf{R})$  over all sub-rectangles, according to the rules discussed earlier. This procedure generates a *non-decreasing* sequence for the lower bound. An *non-increasing* sequence for the upper bound is derived by solving locally the nonconvex problem  $(\mathbf{P})$  and selecting it to be the minimum over all the previously recorded upper bounds. Clearly, if the single minimum of  $(\mathbf{R})$  in any sub-rectangle is greater than the current upper bound we can safely ignore this sub-rectangle because the global minimum of  $(\mathbf{P})$  cannot be situated inside it (fathoming step).

Because the maximum separations between nonconvex terms and their respective convex lower bounding functions are bounded and continuous functions of the size of rectangular domain, arbitrarily small  $\epsilon_f$  feasibility and  $\epsilon_c$  convergence tolerances are reached for a finite size partition element.

The basic steps of the proposed global optimization algorithm and its convergence proof to an  $\epsilon$ -global solution are described in [23].

### 3.3. Minimization of the Conformation Energy using $\alpha\mathbf{BB}$

The deterministic branch and bound type global optimization algorithm  $\alpha\mathbf{BB}$  just described will be utilized so as to bracket the global minimum solution by constructing converging lower and upper bounds. These bounds are successively refined by iteratively partitioning the initial feasible region into many subregions as was pre-

viously described. Upper bounds to the global minimum can be obtained by local minimizations of  $U$ . Lower bounds are obtained by minimizing a convex function  $L$  which is always less than the original nonconvex function  $U$ . This function  $L$  can be constructed by augmenting  $U$  through the addition of a convex separable quadratic term for each dihedral angle.

$$\begin{aligned}
 L = U + \{ & \sum_{i=1}^{N_{RES}} \alpha_{\phi,i} (\phi_i^L - \phi_i) (\phi_i^U - \phi_i) + \\
 & \sum_{i=1}^{N_{RES}} \alpha_{\psi,i} (\psi_i^L - \psi_i) (\psi_i^U - \psi_i) + \\
 & \sum_{i=1}^{N_{RES}} \alpha_{\omega,i} (\omega_i^L - \omega_i) (\omega_i^U - \omega_i) + \\
 & \sum_{i=1}^{N_{RES}} \sum_{k=1}^{K^i} \alpha_{\chi,i,k} (\chi_i^{k,L} - \chi_i^k) (\chi_i^{k,U} - \chi_i^k) + \\
 & \sum_{j=1}^{J^N} \alpha_{j,\theta^N} (\theta_j^{N,L} - \theta_j^N) (\theta_j^{N,U} - \theta_j^N) + \\
 & \sum_{j=1}^{J^C} \alpha_{j,\theta^C} (\theta_j^{C,L} - \theta_j^C) (\theta_j^{C,U} - \theta_j^C) \} \quad (2)
 \end{aligned}$$

Note that  $\phi_i^L, \psi_i^L, \omega_i^L, \chi_i^{k,L}, \theta_j^{N,L}, \theta_j^{C,L}$  and  $\phi_i^U, \psi_i^U, \omega_i^U, \chi_i^{k,U}, \theta_j^{N,U}, \theta_j^{C,U}$  are lower and upper bounds respectively on the dihedral angles  $\phi_i, \psi_i, \omega_i, \chi_i^k, \theta_j^N, \theta_j^C$ . The various  $\alpha$  parameters are the ones defined in Section (3.1).  $\alpha_{BB}$  has been interfaced with ECEPP/3 so as to provide a detailed model of the conformational energy.

The computational requirement of  $\alpha_{BB}$  is proportional to the number of variables on which branching will occur. As a result we should judiciously choose those variables. Qualitatively, a variable on which branching occurs is a variable that participates in a non-convex term and furthermore is expected to greatly influence the location of the global minimum. The recent work of [1, 5, 6] provides a number of key principles so as to identify such important variables, and presents the merits of these principles based on extensive computational studies. It has to be pointed out that should one decide not to branch on a specific variable, the resulting underestimator is still a valid one and properly underestimates the global minimum.

While addressing the problem of protein folding, there exists ample evidence that the most important variables, that is, the ones on which the global minimum is expected to more sensitive to, are the back-bone dihedral angles. Based on this observation we treat the back-bone dihedral angles as the *global* variables on which branching is being performed. The dihedral angles associated with side chains, as

well as the  $\omega$ 's are treated as *local* variables, that is, no branching is being performed. Finally, the user is free to decide as to whether a third set can be defined. This sets include the *fixed* variables, that is, variables which will have constant values. Obviously, the partitioning of the variable set is selected by the user, and the user can specify which dihedral angles belong to which category. The partitioning just described simply defines one such possible instance.

#### 4. Analysis of Oligopeptides

The  $\alpha$ BB coupled with ECEPP/3 has successfully addressed the calculation of the global minimum of single-residue and di-peptides. Various computational results were presented in [25]. The mathematical problem addressed was of the form of (1), whereby the dihedral angles were allowed to vary in the interval  $[-\pi, \pi]$ . In other words no prior knowledge regarding possible ranges within which the dihedral angles could vary was used.

Addressing the conformation of oligopeptides it became apparent that the problem requires more intensive computational effort. The primary reason for this is the quality of the generated lower bounds on the global minimum. In order to address the problem of oligopeptides successfully a meaningful reduction on the bounds of the variables is proposed. These should be such that a substantial size reduction is induced, thus improving the quality of the lower bounds, but also the reduced domains should not exclude regions that contain the global minimum conformation. Having tighter bounds on the global variables results in much tighter lower bounds, provided by the solution of (2), thus improving the computational efficiency of the method.

##### 4.1. Distribution of Dihedral Angle Values

Using hard-sphere models of the atoms and fixed geometries of the bonds, Ramachandran and colleagues, [11], derived regions in terms of the allowed values of the  $(\phi, \psi)$  dihedral angles. The key result of their calculation was that for every naturally occurring amino-acid the structure of these regions remain almost identical.

Similar results were obtained later on, [45], when the distribution of the  $(\phi, \psi)$  angles were recorded for configurations that correspond to low conformational energies based on empirical potential functions (ECEPP/3). In the work of [45], all the dihedral angles were kept constant except the  $(\phi, \psi)$  angles.

Our approach is along these lines and Figure 2 presents the  $(\phi, \psi)$  value distribution for almost 20,000 minimum energy configurations of the 20 naturally occurring amino acids. It should be pointed out, that in our computations none of the dihedral angles are assumed to be fixed, in fact all  $\phi$ ,  $\psi$ ,  $\omega$  and  $\chi$ 's are treated as variables. A total of 1,000 local minima per amino-acid are plotted. Obviously, there is fair degree of repetition and furthermore, a large number of them do not

correspond to physically favorable configurations, although these might be mathematical minima. A substantial number of allowable configurations, corresponding to local minima have been plotted. It is important to emphasize the consistency of the results based on a) the hard sphere model, b) the lower energy conformations, and c) conformations corresponding to local minima. It is clear that regions of very high density can be identified which clearly suggests that the dihedral angles do not assume arbitrary values. The important question that one has to address is whether patterns based on isolated single-residue data could be used to derive bounds on the dihedral angles when these residues are part of a larger molecule.

It was suggested in [18] to analyze the dihedral angle distribution based on the values they assume on polypeptides whose native configuration is known via experimental data. The resulting  $(\phi, \psi)$  maps for the participating residues revealed patterns very similar to the ones obtained when the residues were considered to be isolated. The main point the authors in [18] conveyed is the fact that extrapolating based on the single-residue approach may fail to consider conformation-dependent interactions between residues. Therefore, one should attempt to identify the dihedral angle distribution of specific patterns.

In this direction, we performed an analysis of 95 proteins from the Brookhaven X-ray data bank, [9]. The scope was to derive values for all dihedral angles of the participating naturally occurring amino acids based on experimental data. Our goal was to identify the existence of specific patterns when these residues are a part of a large molecule and based on these observations to detect whether meaningful (i.e., ones that do not exclude global solutions) patterns could be identified. Figure 3 depicts the  $(\phi, \psi)$  map for all the alanine residues identified in the data bank. A total of 1,577 occurrences were recorded. It is interesting to note what happens when one considers the angle distribution of modules composed of more than one alanine residue. Figure 4 depicts the distribution of the *ala-ala* patterns. Clearly, the  $(\phi, \psi)$  become less scattered, and if one considers triplets, that is, *ala-ala-ala* as in Figure 5, the distribution becomes even more focused. Should this evidence be considered universal three key observations can be drawn:

1. All four approaches just described fully agree on the qualitative characteristics of the dihedral angle distributions. The implication of this result is the fact that an underlying symmetry exists and it was properly predicted based on our computational experiments.
2. It appears as if the single-residue approach defines a superset of the possible values that the dihedral angles can assume. Therefore, one should not expect to miss the characteristics of a particular residue when it is considered as a part of an oligo/poly-peptide. Therefore, generalizations based on single-residue data appear to be safe.
3. Based on the results obtained from the 95 proteins of the Brookhaven X-ray data bank considering combinations of single residues appears to be substantially reducing the range of possible values for the dihedral angles. Note that a substantially smaller sample is being used. Only 168 occurrences of *ala-ala* were

observed and 24 occurrences of *ala-ala-ala*. Interestingly enough though, one should observe that the distribution of the values for the *ala-ala-ala* pattern seems to point towards the region that defines the right-handed  $\alpha$  helix which is presumably the configuration of choice for the poly(L-alanine) macromolecules, [38]. In other words, based on the patterns that were identified through the analysis of the 95 proteins, meaningful tight bounds for the poly(L-alanine) have been detected. It should be pointed out however that one should never neglect the risk that the substantial reduction of the  $(\phi, \psi)$  distribution essentially postulates the structure of the poly-peptide. If we consider to focus on the region identified in Figure 5, then we are essentially postulating that the solution can only be a right-handed  $\alpha$ -helix. Such generalizations are extremely dangerous and the risk of missing important information in regard to the global minimum is obvious.

The computational ramifications of these results are very important. One should think in terms of the high dimensionality of the problem in order to identify that if the domain of each variable is halved, for example, the overall domain, for the  $n$ -dimensional case is being reduced by a factor of  $\frac{1}{2^n}$ . We believe that very important information can be derived based on the existence of patterns. Results based on single-residue and multiple-residue patterns will be presented, that will show the viability of the proposed approach.

## 4.2. Partitioning the Search Space

In this section, we will present the procedure for reducing the size of the search domain without excluding the regions of interest, that is, the regions where the global minimum conformation may lie. It will be argued that the range of values for the various dihedral angles,  $\phi, \psi, \omega, \chi$ , is not arbitrary for each independent naturally occurring amino-acid and that distribution patterns can be derived if one analyzes how these values are distributed in a large number of polypeptides.

These results are being presented, in the form of histograms denoting the frequency of occurrence, thus allowing us to identify, per residue and per dihedral angle of that residue, how these values are being distributed. Based on the algorithm used for determining these values the last dihedral angle of the side for each residue, that is a member of a large polypeptide, can not be determined and is therefore treated as a free angle with an arbitrary value. Based on the single residue distributions, correlations will be derived that define regions in a higher space that contain the maximum number of occurrences of these dihedral angles. Results regarding the native conformation of four oligopeptides will be presented, and we discuss the dihedral angle distribution of the amino-acids composing these oligopeptides, namely *tyrosine, glycine, phenylalanine, methionine, leucine, alanine, and proline*.

*tyrosine* : The dihedral angles of interest are  $\phi$ ,  $\psi$ ,  $\omega$ ,  $\chi_1$ , and  $\chi_2$  <sup>1</sup>. Figures (6-10) depict the distribution of the tyrosine angles in a total of 881 occurrences in various polypeptide molecules.

*glycine* : The dihedral angles of interest are  $\phi$ ,  $\psi$ ,  $\omega$  Figures (11-13) depict the distribution of the glycine angles in a total of 1529 occurrences in various polypeptide molecules.

*phenylalanine* : The dihedral angles of interest are  $\phi$ ,  $\psi$ ,  $\omega$ , and  $\chi_1$  <sup>2</sup>. Figures (14-17) depict the distribution of the tyrosine angles in a total of 861 occurrences in various polypeptide molecules.

*methionine* : The dihedral angles of interest are  $\phi$ ,  $\psi$ ,  $\omega$ ,  $\chi_1$ ,  $\chi_2$ , and  $\chi_3$  <sup>3</sup>. Figures (18-23) depict the distribution of the tyrosine angles in a total of 327 occurrences in various polypeptide molecules.

*leucine* : The dihedral angles of interest are  $\phi$ ,  $\psi$ ,  $\omega$ ,  $\chi_1$ ,  $\chi_2$ , <sup>4</sup>. Figures (24-28) depict the distribution of the tyrosine angles in a total of 1543 occurrences in various polypeptide molecules.

*alanine* : The dihedral angles of interest are  $\phi$ ,  $\psi$ ,  $\omega$  <sup>5</sup>. Figures (29-31) clearly define two regions of interest. 1577 occurrences of *alanine* were detected.

*proline* : The dihedral angles of interest are  $\phi$ ,  $\psi$ ,  $\omega$  and shown in Figures (32-34). 38 occurrence of proline were identified.

The following important observations have to be made based on these results:

1. It is clear from these figures that a tremendous symmetry exists and that the values of all dihedral angles are by no means random.
2.  $\omega$  seems to be restricted in a range of values centered around  $\pi$ .
3. The side-chain dihedral angles (i.e., the  $\chi$ 's) seem to be distributed, in a structured way, throughout a wide range of values. Most previous attempts neglect the importance of the  $\chi$  values. Although in large molecules the primary dihedral angles responsible for identifying the correct folded state are the backbone dihedral angles, it seems as if side-chain dihedral angles are expected to play an important role in determining the energetically most stable configuration.

Based on the distributions observed in Figures (6 - 34), reduced domains for the dihedral angles of various amino acids can be identified. By observing Figure (11), for example, one can deduce that with respect to the  $\phi$  angle of *glycine* we can identify two distinct regions. The first in the interval  $[-180, -30]$ , and the other in the interval  $[30, 180]$ . On the other hand, with respect to the  $\psi$  values, Figure (12), we can observe that the values are distributed throughout the range  $[-180, 180]$ , and therefore we select to define the two regions  $[-180, 0]$  and  $[0, 180]$ . By using similar arguments we end up in the regions which are defined in Table 1. Note that Table 1 consists of a number of regions based on the partitioning of the dihedral angles.

Table 1. Bounds on dihedral angles. 1: bounds based on the -ala-ala-ala- pattern, 2: bounds based on the -ala- pattern, 3: The  $\phi$  value for the down packing of proline that ECEPP3 uses is -68.8

	$\phi$	$\psi$	$\omega$	$\chi_1$	$\chi_2$	$\chi_3$	$\chi_4$
tyr	-180,0	-75,50 50,175	160,200	-180,180	-180,180	-180,180	
gly	-180,-30 30,180	-180,0 0,180	160,200				
phe	-180,-50	-75,50 50,175	160,200	-180,180	-180,180		
met	-180,-50	-75,50 50,175	160,200	-180,180	-180,180	-180,180	-180,180
leu	-180,-50	-75,50 50,175	160,200	-180,180	-180,180	-180,180	-180,180
ala <sup>1</sup>	-150,-50	-100,0	160, 200	-180, 180			
ala <sup>2</sup>	-180,-50 180, 50	-75,-25 50,175	160, 200 160, 200	-180, 180 -180, 180			
pro	68.8 <sup>3</sup>	-75,0 150,200	160, 200 160, 200				

It should be noted finally that significant reductions can be derived only for the backbone dihedral angles. The side chain angles,  $\chi$ 's, are distributed in a structured yet not simply described, using discrete domains, sub-regions. If one defines the sub-intervals of the  $\chi$  angles using similar arguments the initial number of sub-regions to be examined will increase, as will be explained in the following section.

### 4.3. A Novel Scheme for Partitioning of Domains in $\alpha\text{BB}$

Based on the above, for every residue and every dihedral angle in that residue, a set of reduced domains for the backbone dihedral angles can be defined. Consider for example the  $\psi$  angle of *tyrosine*, that is,  $i = \text{tyrosine}$ . For the dihedral angles,  $\psi$ , we have identified two domains,  $\Psi_{i\psi}$ ,  $i_\psi = 1, \dots, N_{i\psi}$  with  $N_{i\psi}$  the number of domains for the  $\psi$  angle of the  $i$  residue, and  $\Psi_{i\psi} = \{(-75, 50), (50, 175)\}$ . Similarly, these quantities are being defined for all other dihedral angles.

In this case the original problem as defined in (1) is now defined as follows :

$$\begin{aligned} \min \quad & U(\phi_i, \psi_i, \omega_i, \chi_i^k, \theta_j^N, \theta_j^C) \\ \text{subject to} \quad & \phi_i \in \Phi_{i\phi}, \quad i_\phi = 1, \dots, N_{i\phi}, \quad i = 1, \dots, N_{RES} \\ & \psi_i \in \Psi_{i\psi}, \quad i_\psi = 1, \dots, N_{i\psi}, \quad i = 1, \dots, N_{RES} \end{aligned}$$

Table 2. Domains for Met-enkephalin

	$ \Phi $	$ \Psi $	$ \Omega $	$ X_1 $	$ X_2 $	$ X_3 $	$ X_4 $	$N_i$
tyr	1	2	1	1	1	1		2
gly	2	2	1					4
gly	2	2	1					4
phe	1	2	1	1	1			2
met	1	2	1	1	1	1	1	2
N								128

$$\begin{aligned} \omega_i &\in \Omega_{i\omega}, \quad i_\omega = 1, \dots, N_{i\omega}, \quad i = 1, \dots, N_{RES} \\ X_i^k &\in X_{iX}^k, \quad i_X = 1, \dots, N_{iX}^k, \quad i = 1, \dots, N_{RES}, \quad k = 1, \dots, K(\S) \\ \theta_j^N &\in \Theta_{j\theta^N}, \quad j_\theta^N = 1, \dots, N_{\theta^N}^j, \quad j = 1, \dots, J_N \\ \theta_j^C &\in \Theta_{j\theta^C}, \quad j_\theta^C = 1, \dots, N_{\theta^C}^j, \quad j = 1, \dots, J_C \end{aligned}$$

Clearly, the allowable dihedral angles for each residue should belong to one of the following domains:

$$\begin{aligned} D_{ji} &= \Phi_{i_\phi} \times \Psi_{i_\psi} \times \Omega_{i_\omega} \times X_{i_X}^k \\ ji &= 1, \dots, N_i \\ N_i &= |\Phi_{i_\phi}| |\Psi_{i_\psi}| |\Omega_{i_\omega}| |X_{i_X}^k| \end{aligned}$$

where  $M_i$  is the number of dihedral angles with the maximum number of domains. Based on the above definitions and the partitioning of the search space, the total number of initial domains is given by :

$$N = \left( \prod_{i=1}^{N_{RES}} N_i \right) |\Theta_{j_\theta^N}| |\Theta_{j_\theta^C}|$$

Note that “ $|\bullet|$ ” denotes the cardinality of the set. These domains correspond to the cartesian products of all the sub-domains  $D_{ij}$ ,  $ji = 1, \dots, N_i$ .

As an illustrative example, let us consider the molecule of Met-enkephalin. There are a total of 5 residues,  $N_{RES} = 5$ . Table 2 provides the values for  $N_i$  and  $N$  for the amino-acids participating in Met-enkephalin.

The aforementioned procedure has now defined for Met-enkephalin a finite number of domains  $D_i \subset [-\pi, \pi]^{NDA}$ ,  $i = 1, \dots, 128$ , where  $NDA$  is the total number of dihedral angles which constitute the starting domains for  $\alpha BB$ . Furthermore, this procedure can be nicely integrated within a distributed framework for  $\alpha BB$  that is under development, [7]. From an implementation point of view, instead of initializing  $\alpha BB$  with a single domain, that is subsequently partitioned based on the

Table 3. Dihedral angles at the global minimum potential energy conformation of Ac-Ala<sub>4</sub>-Pro-NHMe.

	$\phi$	$\psi$	$\omega$	$\chi_1$
Ac	-179.997	-178.210		
Ala	-71.102	-27.158	178.745	-178.378
Ala	-70.349	-35.997	183.369	61.303
Ala	-80.986	-36.950	184.664	-58.108
Ala	-133.711	71.251	176.872	179.263
Pro	-68.8	-24.235	179.832	
NHMe	59.915			

steps of the algorithm,  $N$  domains are being created which are all assigned appropriate values for their corresponding lower bounds. In other words,  $\alpha\mathcal{B}\mathcal{B}$  is being initialized with a set of consistent domains provided by the domain partitioning studies that can provide tight initial lower bounds. Note that had we partitioned further the domains of the  $\chi$  angles, a substantial increase in the number of initial domains would have occurred. In order to avoid that we treat the  $\chi$  angles as local variables, in the context discussed earlier, based on the assumption that the effect of the side chain will be not as important as the effect of the backbone structure in the prediction of the correct conformation.

## 5. Computational Studies

The proposed approach has been tested on a number of oligopeptide potential energy minimization instances. The selected relative convergence tolerance is  $10^{-2}$  and the computational requirements reported in seconds are on an HP-730 workstation. In all cases we treat as global variables the  $\phi$ 's and  $\psi$ 's and as local all the remaining, (i.e.,  $\omega$ 's and  $\chi$ 's) where appropriate. In all of our computational studies the value of the  $\alpha$  parameter was set to 3.5.

### 5.1. N-Acetyl-N<sup>2</sup>-methylamide of Ala<sub>4</sub>-Pro

The first example is concerned with the lowest potential energy structure of Ac-Ala<sub>4</sub>-Pro-NHMe involving 21 dihedral angles. This problem was first proposed in [34] for evaluating the capability of ECEPP/3 to correctly describe the energetics of the atomic interactions. The global minimum potential energy conformation of Ac-Ala<sub>4</sub>-Pro-NHMe is characterized by a potential energy value of  $-18.91$  kcal/mole. Note that this potential energy value is slightly better than the one reported in [34] ( $-18.82$  kcal/mole). Table 3 summarizes the values of the 21 dihedral angles at the global minimum solutions. A plot of this conformation is shown in Figure 35.

In the study of Ala<sub>4</sub>-Pro results for both the single and multiple residue patterns were considered. An analysis of the pattern created while studying the complex -ala-ala-ala- reveals a very tight domain for the  $\phi$  and  $\psi$  angles. Therefore, we use these bounds for the backbone angles of the first three *alanine* residues. In regard to the fourth *alanine* residue, that is, the one directly connected to the *proline* molecule, we allowed the dihedral angles to vary according to the domains identified in Table 1. This arrangement results in an initial partitioning of 4 domains. Convergence to the optimal solution was achieved after 422 iterations and 2,816 s. It should be pointed out that a strong local minimum exists within only about 0.3 kcal/mol higher energy whose vector of dihedral angles is practically identical to the global minimum with only the exception of the  $\psi$  angle of proline. The global minimum corresponds to a distorted right-handed  $\alpha$  helix, and the source of the distortion is the influence of the proline.

## 5.2. Met-enkephalin

This example illustrates the application of the proposed approach on a benchmark molecular conformation problem. It involves the identification of the global minimum total potential energy conformation of the penta-peptide molecule Met-enkephalin. Met-enkephalin (H-Tyr-Gly-Gly-Phe-Met-OH), is an endogenous opioid linear penta-peptide found in the human brain, pituitary, and peripheral tissues. Its biological function is related with the endogenous response to pain and a large variety of physiological processes. Met-enkephalin consists of 75 atoms and it involves 24 independent dihedral angles, giving rise to a very complex conformational space involving a plethora of local minima which are estimated in the order of  $10^{11}$  ([20]). Met-enkephalin represents a very challenging conformation study because of (i) the large number of local minima which make the energy hypersurface rather bumpy, (ii) the existence of strong local minima that most local optimization algorithms terminate, (iii) the location of these strong local minima is not necessarily close to the global minimum even though there are strong local minima within 1.0 kcal/mole, (iv) the very special initialization procedures that need to be devised so as to reach the best solution, (v) the failure of any local optimization method if random initial points are considered, and (vi) the reported failure of local optimization methods to locate the global solution even though the starting point is very close to the global minimum.

The global minimum potential energy conformation of Met-enkephalin is shown in Figure 36. The values of the 24 dihedral angles at the global minimum conformation are given in Table 4.

As one can observe from Figure 36 the global minimum configuration exhibits a bend along the  $N - C'$  peptidic bond of *Gly*<sup>3</sup> and *Phe*<sup>4</sup>, and it represents a type II'  $\beta$ -bend, [43], consistent with observations made earlier, [14].

Based on the analysis presented earlier, for all residues the  $\psi$  domain is partitioned in two sub-domains and furthermore the  $\phi$  domain of the *glycine* residue is also partitioned in to two domains. Table 2 presents the corresponding partitioned

Table 4. Dihedral angles at the global minimum potential energy conformation of Met-enkephalin.

	$\phi$	$\psi$	$\omega$	$\chi_1$	$\chi_2$	$\chi_3$	$\chi_4$
Tyr	-83.457	155.787	-177.159	-173.205	-100.544	13.636	
Gly	-154.298	86.001	168.507				
Gly	82.964	-75.115	-169.939				
Phe	-136.880	19.079	-174.065	58.843	-85.543		
Met	-163.465	160.944	-179.794	52.884	175.268	-179.851	61.421

domains. This results in a partitioning of 128 domains that are considered in the initialization step. Based on the partitioning according to the single residue data we define a set of problems, on disjoint domains, that are to be solved. In other words, we consider that the first 7 levels of the branch and bound tree have collapsed into one level, the initial one, composed of 128 nodes, where each node represents one of these domains. Unlike other approaches, no assumptions are made regarding the values of the  $\chi$  dihedral angles and they are allowed to vary in the entire interval  $[-\pi, \pi]$ . In terms of the computational requirements it takes  $\alpha$ BB 977 iterations, and 4,669 seconds in order to converge within the required tolerance. It should be noted that the aforementioned CPU time includes the determination of about 100 low energy conformations close to the global minimum of  $-11.707$ . At each iteration of the  $\alpha$ BB method, the solution of an upper bounding local optimization problem and two convex lower bounding problems are needed. For all three problems the nonlinear solver requires 2-4 CPU seconds per iteration and within this time there are on the average 400 calls to ECEPP/3. At each call to ECEPP/3, the input is a set of values of the dihedral angles. Then, ECEPP/3 transforms the internal coordinates into cartesian coordinates, performs the function and gradient evaluation in the cartesian space and transforms them back into the internal coordinates to be used by the local optimization solver. Each call to ECEPP/3 requires 0.005-0.01 CPU seconds, and hence there exists a rather fast calculation of the functions and gradients which in turn require rapid calculation of the nonbonded distances from the internal coordinates.

Over the years, Met-enkephalin has received a fair amount of attention and there exists an abundance of computational results that could be used for comparison purposes. Based on the available computational experience the following points can be made:

### (i) Comparison of $\alpha$ BB with other approaches

At the outset of such a comparison, it should be emphasized that the  $\alpha$ BB (i) offers theoretical guarantee of determining an  $\epsilon$ -global minimum solution for twice-differentiable optimization problems with analytical functions, (ii) identifies valid upper and lower bounds on the global solution and (iii) identifies local

optima which are close to the global minimum one during the global optimum search.

The primary objective in such a comparison of  $\alpha$ BB with other more standard procedures is to investigate the computational effort required by the  $\alpha$ BB in contrast to the other approaches. Since Met-enkephalin has been studied extensively in the literature with a variety of methods, reported results for several different methods will be presented. Note that due to the difficulties in the search for the global minimum structure of Met-enkephalin there exist a few studies that have treated all 24 dihedral angles as variables, while most of the studies have considered either 10 dihedral angles (i.e.,  $\phi$ , and  $\psi$ ) or 19 dihedral angles (i.e.,  $\phi$ ,  $\psi$ , and  $\chi$ ).

Table 5 summarizes the computational requirements for addressing the same problem with a variety of various approaches. Based on the above results we can clearly see that  $\alpha$ BB compares favorably with most of the widely used methods, and furthermore it provides additional information regarding not only bounds on the global minimum total potential energy, but also on bounds on the values of the dihedral angles. Note also that  $\alpha$ BB unlike simulated annealing and/or Monte Carlo approaches, is a domain based method and not a point based method. Therefore, one identifies regions of the search space rather than singletons.

## (ii) Comparison with local optimization methods

A case against using global optimization methods is often made based on the computational requirements and is further suggested that an equivalent amount of local runs would have produced similar results. In order to test such a statement, a number of local runs equivalent to the number of times the upper bound problem was solved with  $\alpha$ BB were performed. Based on the results just presented  $\alpha$ BB required about 1,000 iterations implying that the non-convex problem was solved about 1,000 times so as to provide valid upper bounds for  $\alpha$ BB. An equivalent amount of local runs were performed from randomly generated starting points and the results were recorded.

The configuration with the lowest potential energy that is obtained has a value of  $E = -8.002 \text{ kcal/mole}$  which is over 30% higher in energy than the global minimum potential energy conformation with a potential energy value of  $E = -11.707 \text{ kcal/mole}$ . These results clearly suggest the inadequacy of multistart-like local optimization techniques to address such a complicated problem.

It should also be pointed out that [32] provided a direct comparison of the Simulated Annealing, SA, method with the Monte Carlo Minimization, MCM, method for Met-enkephalin. The conclusions drawn from this comparison are : (i) the SA converges to a lower energy structure faster than the MCM method but the SA does not converge to the global minimum in any of the 24 randomly chosen initial conformations whereas the MCM does converge; (ii) the energy difference between the minima reached by the two methods is typically 5-15

Table 5. Computational results on Met-enkephalin

Method	$N_{var}$	CPU	Computer
Monte Carlo Minimization [20]	19	2-3 hrs	IBM 3090
	24	10 hrs	IBM 3090
Electrostatically Driven Monte Carlo [39]	19	2-3 hrs	IBM 3090
	24	10 hrs	IBM 3090
Diffusion Equation [17]	19	20 min	IBM 3090
Self-Consistent Multitorisional Field [36]	10	100 min	IBM 3090
Multicanonical Simulated Annealing [15]	19	6. hrs	IBM RS600
Simulated Annealing [30], 1991)	24	2.5 hrs	Apollo DN1000
Threshold Simulated Annealing [31]	24	1.5 hrs	Apollo DN1000
Simulated Annealing with Monte Carlo Minimization [13]	24	2 hrs	CRAY X-MP
Simulated Annealing with Monte Carlo Minimization [42]	24	1.2 hrs	CRAY-2S, 4 processors
Monte Carlo Minimization vs. Simulated Annealing [32]	24	1.5-4 hrs	IBM 3090
$\alpha$ BB	24	1.3 hrs	HP-730

kcal/mole in favor of MCM; (iii) even though the SA is followed by local energy minimization the global optimum solution cannot be reached (i.e., there was an energy decrease of 2.3 kcal/mole but it was far from the global optimum by 6 kcal/mole), and the RMS deviation was not improved; (iv) even though instead of using the final conformation reached at the end of a given temperature interval, the minimum energy conformation encountered during that interval was used as the starting point for the next (lower) temperature interval, the global minimum structure could not be reached (i.e., there was a decrease of 2 kcal/mole in energy).

### (iii) Low energy conformers

One of the primary advantages of  $\alpha$ BB, is that several low energy conformations are also provided during the global optimization search. Such information is very important. Table 6 summarizes five low energy conformers of Met-enkephalin along with their RMS deviation values, in cartesian coordinates, when compared to the global optimum structure. Note that these are strong local minima and if we initiate a search from these structures we cannot obtain the global minimum conformation. Also note that this occurs even though the

Table 6. Low Energy Conformers with RMS deviations computed in the cartesian space.

Conf. No.	$U$	RMS deviation
1	-11.707	Global Minimum
2	-11.696	0.203905
3	-11.164	1.132507
4	-11.117	1.196889
5	-10.781	3.175264

Table 7. Dihedral angles for Leu-enkephalin.

	$\phi$	$\psi$	$\omega$	$\chi_1$	$\chi_2$	$\chi_3$	$\chi_4$
Tyr	102.135	139.309	-166.873	-163.758	79.920	-154.795	
Gly	00.000	60.323	179.446				
Gly	69.574	-85.997	174.502				
Phe	-101.036	-24.937	-172.496	73.661	86.759		
Leu	-81.092	131.465	177.678	-178.432	65.431	-67.792	-180.500

RMS deviation values of these local optimal do not differ substantially from the global optimum structure. In fact, all these five conformations exhibit similar hairpin structures as the global minimum one.

### 5.3. Leu-enkephalin

This peptide is also an endogenous pentapeptide much like met-enkephalin with the only difference being the fact that *methionine* has been replaced by *leucine*. Figure 37 shows the obtained minimum energy conformation of the isolated molecule which is once again a type II'  $\beta$ -bend around the  $Cly^3 - Phe^4$  backbone region, [14], similar to the one observed for Met-enkephalin. Using a similar approach for initiating  $\alpha$ BB, it takes 1027 iterations and 5,209 s. in order to converge within the required tolerance, with a total potential energy of  $E = -9.332 \text{ kcal/mole}$ . The values of the 24 dihedral angles are given in Table 7.

### 5.4. Decaglycine

This example concerns the minimization of the total potential energy of a larger oligopeptide, namely *decaglycine*. Decaglycine consists of 30 dihedral angles that are to be optimized and was studied very thoroughly by [40] using the EDMC method coupled with ECEPP/2. The number of local minima is enormous and

these metastable structures correspond to partially right-hand and left-hand  $\alpha$ -helices. Based on the computational experience available for this oligopeptide a unique starting region was defined whose bounds were set to be  $\pm 60$  degrees around the best solution reported in [40]. It took  $\alpha$ BB a total of 102 iterations and about 14,000 seconds to identify the configuration with  $E = -11.642$ . The generated helix is shown in Figure 38. Clearly, the computational requirements is an indication of the difficulty of the problem. It should be pointed out that *glycine* is flexible and can assume a large number of configurations without introducing any substantial amount of steric interactions. Along those lines, one should consider patterns of more than one residues so as to reduce the computational effort.

## 5.5. Computational Complexity

Based on the computational results presented in this work, as well as on previous results [25] that we have obtained within the framework of  $\alpha$ BB, the behavior of this approach is considered to be fairly consistent. The CPU requirements are in the order of  $n^3$  and the study of Met-enkephalin showed that they compare favorably with Monte Carlo and Simulated Annealing methods.

## 6. Conclusions

In this paper we presented a systematic procedure for identifying (i) the global minimum energy conformation of oligopeptides, (ii) upper and lower bounds on the global minimum energy, and (iii) several low energy conformers. The procedure is based on the deterministic global optimization algorithm  $\alpha$ BB and uses the ECEPP/3 potential energy model in a unified framework. Distribution patterns of the various dihedral angles of the naturally occurring amino acids were determined based on an analysis of a large number of proteins whose native configuration was known experimentally. Analysis of single residue data allowed the identification of domains in the  $(\phi, \psi)$  space with high probabilities. Combinations of these domains defined the starting points for  $\alpha$ BB. The computational efficiency of the method was demonstrated by identifying the global minimum energy conformation of the three penta-peptides, namely Met-enkephalin, Leu-enkephalin, and Ac-Ala<sub>4</sub>-Pro-NHMe, as well as the deca-peptide decaglycine. Extensions of this work into analyzing pattern formation of multiple residue building blocks that would allow to address poly-peptides is in progress. The proposed approach provides a natural decomposition of the search domain and is easily parallelizable.

It should be pointed out that the global optimization approach  $\alpha$ BB offers theoretical guarantee of attaining a global optimum solution for general twice-differentiable nonlinear optimization problems in which the objective function and constraints are provided analytically. In this work we employ fixed values of the  $\alpha$  parameter and use the force field ECEPP/3 for the function and gradient evaluations and not an analytical expression for the objective function even though an analytical mapping

exists. As a result the issue of providing theoretical guarantee for the combined ECEPP/3 -  $\alpha$ BB approach still remains. Work in this direction is based on rigorous bounds on the minimum eigenvalues [2] and is currently pursued.

### Acknowledgements

Financial support from the National Science Foundation, the Air Force Office of Scientific Research, the National Institute of Health, as well as Exxon Co., and Mobil Corporation is acknowledged.

### Notes

1.  $\chi_3$  can not be determined based on the information provided by the PDB-files
2.  $\chi_2$  can not be determined based on the information provided by the PDB-files
3.  $\chi_4$  can not be determined based on the information provided by the PDB-files
4.  $\chi_3$  and  $\chi_4$  can not be determined based on the information provided by the PDB-files
5.  $\chi_1$  can not be determined based on the information provided by the PDB-files

### References

- [1] C. Adjiman, I.P. Androulakis, C.D. Maranas, and C.A. Floudas, A Global Optimization Method,  $\alpha$ BB, for Process Design. *Comp. and Chem. Eng.*, **20**:S419-S424, 1996.
- [2] C. Adjiman and C.A. Floudas, Rigorous Convex Underestimators for General Twice-Differentiable Problems. *J. Global Opt.*, **9**:23-40, 1996.
- [3] N.L. Allinger, Conformational Analysis. MM2: A Hydrocarbon Force Field Utilizing  $V_1$  and  $V_2$  Torsional Terms. *J. Am. Chem. Soc.*, **99**:8127-8134, 1977.
- [4] N.L. Allinger, Y.H. Yuh, and J.-H. Lii, Molecular Mechanics. The MM3 Force Field for Hydrocarbons. *J. Am. Chem. Soc.*, **111**:8551-8582, 1989.
- [5] I.P. Androulakis, C.D. Maranas and C.A. Floudas,  $\alpha$ BB: A Global Optimization Method for General Constrained Nonconvex Problems. *J. Global Opt.*, **7**:337-363, 1995.
- [6] I.P. Androulakis, C.D. Maranas, and C.A. Floudas,  $\alpha$ BB: A Global Optimization Method for General Constrained Nonconvex Problems. Paper presented at the AIChE Annual Meeting, Miami, FL, 1995.
- [7] I.P. Androulakis, C.D. Maranas, and C.A. Floudas, Distributed Branch and Bound Algorithms for Global Optimization. Paper presented at the AIChE Annual Meeting, Miami, FL, 1995.
- [8] C.B. Anfinsen, E. Haber, M. Sela, and F.H. White, The Kinetics of Formation of Native Ribonuclease During Oxidation of the Reduced Polypeptide Chain. *J. Proc. Nat. Acad. Sci. U.S.A.*, **47**:1309-1314, 1961.
- [9] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer Jr, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, the Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. *J. Mol. Biol.*, **12**:535-542, 1977.
- [10] B. Brooks, R. Bruccoleri, B. Olafson, D. States, S. Swaminathan, and M. Karplus, CHARM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comp. Chem.*, **4**:187-217, 1983.

- [11] T.E. Creighton, *Proteins: Structures and Molecular Properties*, W.H. Freeman and Company, New York, 1993.
- [12] P. Dauber-Osguthorpe, V.A. Roberts, D.J. Osguthorpe, J. Wolff, M. Genest and A.T. Hagler, Structure and Energetics of Ligand Bindings to Peptides: Escherichia coli Dihydrofolate Reductase-Trimethoprim, A Drug Receptor System. *Proteins: Struct. Funct. Genet.*, **4**:31-47, 1988.
- [13] B. Freyberg and W.J. Braun, Efficient Search for All Low Energy Conformations of Polypeptides by Monte Carlo Methods. *J. Comp. Chem.*, **12**:1065-1076, 1991.
- [14] L. Glasser and H.A. Scheraga, Calculations on Crystall Packing of a Flexible Molecule, Leu-Enkephalin. *J. Mol. Biol.*, **199**:513-524, 1988.
- [15] U.H. Hansmann and Y. Okamoto, Prediction of Peptide Conformation by Multicanonical Algorithm: New Approach to the Multiple-Minima Problem. *J. Comp. Chem.*, **14**:1333-1338, 1993.
- [16] A.J. Hopfinger, *Conformational Properties of Macromolecules*. Academic Press, New York, NY, 1973.
- [17] J. Kostrowicki and H.A. Scheraga, Application of the Diffusion Equation Method for Global Optimization to Oligopeptides. *J. Phys. Chem.*, **96**:7442-7449, 1992.
- [18] M.H. Lambert and H.A. Scheraga, Payttern Recognition in the Prediction of Protein Structure I. Tripeptide Conformational Probabilities Calculated from the Amino Acid Sequence. *J. Comp. Chem.*, **6**:770-797, 1989.
- [19] M. Levitt, Protein Folding by Restrained Energy Minimization and Molecular Dynamics. *J. Mol. Biol.*, **170**:723-764, 1983.
- [20] Z. Li and H.A. Scheraga, Structure and Free Energy of Complex Thermodynamic Systems. *J. Mol. Struct. (Theochem.)*, **179**:333-352, 1988.
- [21] C.D. Maranas and C.A. Floudas, A Global Optimization Approach for Lennard-Jones Microclusters. *J. Chem. Phys.*, **97**:7667-7678, 1992.
- [22] C.D. Maranas and C.A. Floudas, Global Optimization for Molecular Conformation Problems. *Ann. Oper. Res.*, **42**:85-117, 1993.
- [23] C.D. Maranas and C.A. Floudas, Global Minimum Potential Energy Conformations of Small Molecules. *J. Global Opt.*, **4**:135-170, 1994.
- [24] C.D. Maranas and C.A. Floudas, A Deterministic Global Optimization Approach for Molecular Structure Determination. *J. Chem. Phys.*, **100**:1247-1261, 1994.
- [25] C.D. Maranas, I.P. Androulakis, and C.A. Floudas, A Deterministic Global Optimization Approach for the Protein Folding Problem, *Global Optimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*. P.M. Pardalos, D. Shalaway, and G. Xue (Eds.). DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, **23**:133-150, 1996.
- [26] B.A. Murtagh and M.A. Saunders, *MINOS5.0 Users Guide*. Systems Optimization Laboratory, Dept. of Operations Research, Stanford University, CA., 1983.
- [27] F.A. Momany, L.M. Carruthers, R.F. McGuire, and H.A. Scheraga, Intermolecular Potentials from Crystal Data. III.. Determination of Empirical Potentials and Applications to the Packing Configurations and Lattice Energies in Crystals of Hydrocarbons, Carboxylic Acids, and Amides. *J. Phys. Chem.*, **78**:1595-1620, 1974.
- [28] F.A. Momany, L.M. Carruthers, and H.A. Scheraga, Intermolecular Potentials from Crystal Structures. III. Application of Empirical Potentials to the Packing Configurations and Lattice Energies in Crystals of Amino Acids. *J. Phys. Chem.*, **78**:1621-1630, 1974.
- [29] F.A. Momany, R.F. McGuire, A.W. Burgess, and H.A. Scheraga, Energy Parameters in Polypeptides. VII Geometric Parameters, Partial Atomic Charges, Nonbonded Interactions, Hydrogen Bond Interactions, and Intrinsic Torsional Potentials for the Naturally Occuring Amino Acids. *J. Phys. Chem.*, **79**:2361-2381, 1975.

- [30] L.B. Moralles, R. Garduno-Juarez, and D. Romero, Applications of Simulated Annealing to the Multiple-Minima Problem in Small Peptides. *J. Biomol. Struct. Dynamics*, **8**:721-735, 1991.
- [31] L.B. Moralles, R. Garduno-Juarez, and D. Romero, The Multiple-Minima Problem in Small Peptides Revisited. The Threshold Accepting Approach. *J. Biomol. Struct. Dynamics*, **9**:951-957, 1992.
- [32] A. Nayeem, J. Vila, and H.A. Scheraga, A Comparative Study of the Simulated-Annealing and Monte Carlo-with-Minimization Approaches to the Minimum-Energy Structures of Polypeptides: Met-Enkephalin. *J. Comp. Chem.*, **12**:594-605, 1991.
- [33] G. Némethy, M.S. Pottle, and H.A. Scheraga, Energy Parameters in Polypeptides. 9. Updating of Geometrical Parameters, Nonbonded Interaction, and Hydrogen Bond Interactions for the Naturally Occurring Amino Acids. *J. Phys. Chem.*, **89**:1883-1887, 1983.
- [34] G. Némethy, K.D. Gibson, K.A. Palmer, C.N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, and H.A. Scheraga, Energy Parameters in Polypeptides. 10. Improved Geometrical Parameters and Nonbonded Interactions for Use in the ECEPP/3 Algorithms, with Applications to Proline-Containing Peptides. *J. Phys. Chem.*, **96**:6472-6484, 1992.
- [35] A. Neumaier, Molecular Modeling of Proteins: The Mathematical Prediction of Protein Structure. *SIAM Review*, submitted for publication, 1995.
- [36] K.A. Olszewski, L. Piela, and H.A. Scheraga, Mean Field Theory as a Tool for Intramolecular Conformational Optimization. Tests on Terminally Blocked Alanine and Met-enkephalin. *J. Phys. Chem.*, **96**:4672-4676, 1992.
- [37] P.M. Pardalos, D. Shalloway, and G. Xue (Eds.), *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*, DIMACS series in Discrete Mathematics and Theoretical Computer Science, **23**, Providence, RI, American Mathematical Society, 1996.
- [38] D.R. Ripoll and H.A. Scheraga, On the Multiple-Minima Problem in the Conformational Analysis of Polypeptides. II. An Electrostatically Driven Monte Carlo Method - Tests of Poly(L-Alanine). *Biopolymers*, **27**:1283-1303, 1988.
- [39] D.R. Ripoll and H.A. Scheraga, The Multiple Minima Problem in the Conformational Analysis of Polypeptides. III. An Electrostatically Driven Monte Carlo Method - Tests on Enkephalin. *J. Protein Chem.*, **8**:263-287, 1989.
- [40] D.R. Ripoll, M. Vásquez, and H.A. Scheraga, The Electrostatically Driven Monte Carlo Method: Application to conformational Analysis of Decaglycine. *Biopolymers*, **31**:319-330, 1991.
- [41] H.A. Scheraga, *Reviews in Computational Chemistry*. VCH Publishers, New York, NY, 1992.
- [42] J. Shin and M. Jhon, High Directional Monte Carlo Procedure Coupled with the Temperature and Annealing Method to Obtain the Global Energy Minimum Structure of Polypeptides and Proteins. *Biopolymers*, **31**:177-185, 1991.
- [43] B.L. Sibanda and J.M. Thornton,  $\beta$ -Hairpin families in globular proteins. *Nature*, **316**:170-174, 1985.
- [44] W.F. van Gunsteren and H.J.C. Berendsen, *GROMOS*. Groningen Molecular Simulation, Groningen, The Netherlands, 1987.
- [45] M. Vásquez and H.A. Scheraga, Use of Buildup and Energy-Minimization Procedures to Compute Low-Energy Structures of the Backbone of Enkephalin. *Macromolecules*, **16**:1043-1049, 1983.
- [46] M. Vásquez, G. Némethy, and H.A. Scheraga, Conformation Energy Calculations on Polypeptides and Proteins. *Chem. Rev.*, **94**:2183-2239, 1994.
- [47] S. Weiner, P. Kollmann, D.A. Case, U.C. Singh., C. Ghio, G. Alagona, S. Profeta and P. Weiner, A New Force Field for Molecular mechanical Simulation of Nucleic Acids and Proteins. *J. Am. Chem. Soc.*, **106**:765-784, 1984.

- [48] S. Weiner, P. Kollmann, D. Nguyen, and D. Case, An All Atom Force Field for Simulations of Proteins and Nucleic Acids. *J. Comp. Chem.*, **7**:230-252, 1986.

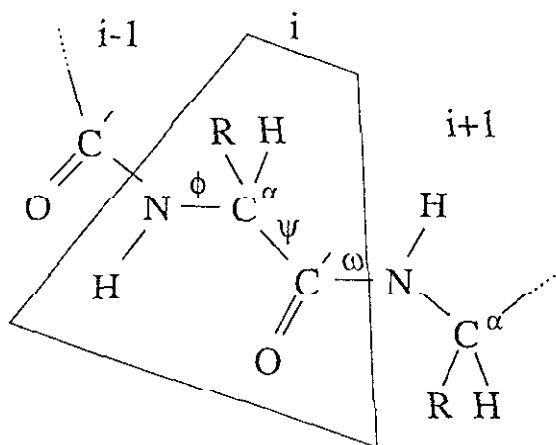
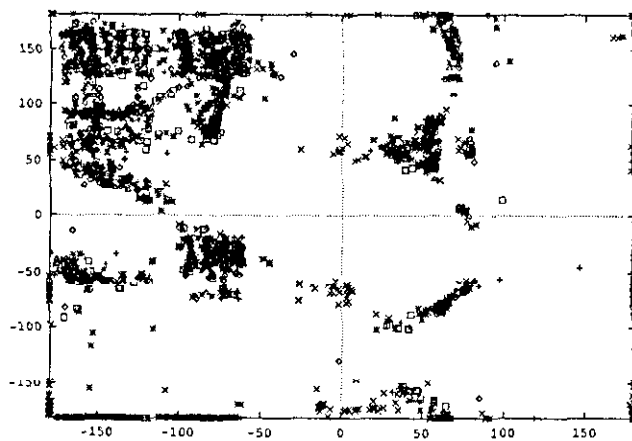


Figure 1: Dihedral angles in a protein

Figure 2:  $(\phi, \psi)$  map denoting backbone conformational states for 20,000 minimum configurations of the naturally occurring amino acids

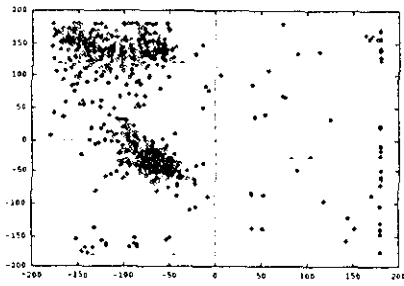


Figure 3:  $(\phi, \psi)$  map for the  $-ala-$  pattern

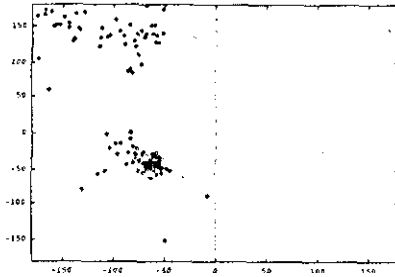


Figure 4:  $(\phi, \psi)$  map for the  $-ala-ala-$  pattern

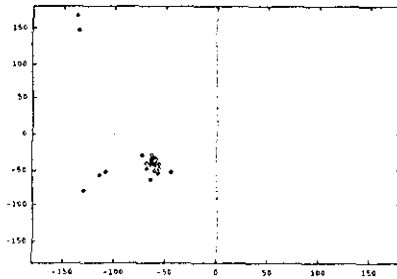


Figure 5:  $(\phi, \psi)$  map for the  $-ala-ala-ala-$  pattern

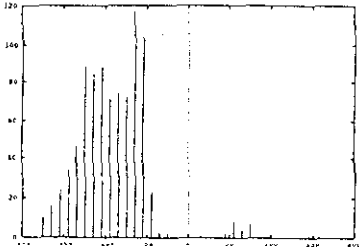


Figure 6:  $\phi$  of tyrosine

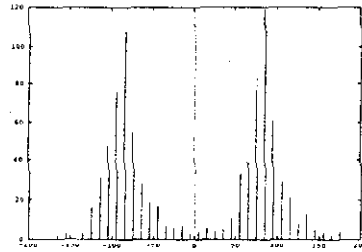


Figure 10:  $\chi_2$  of tyrosine

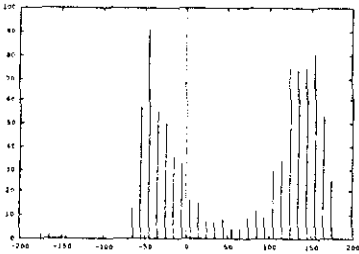


Figure 7:  $\psi$  of tyrosine

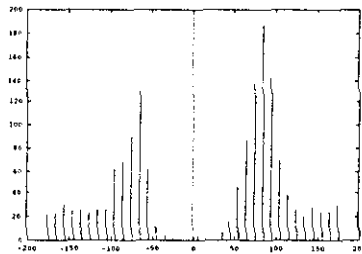


Figure 11:  $\phi$  of glycine

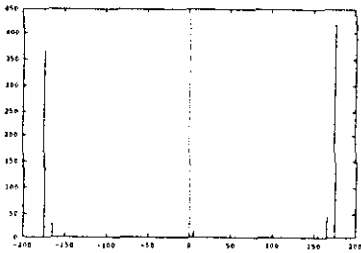


Figure 8:  $\omega$  of tyrosine

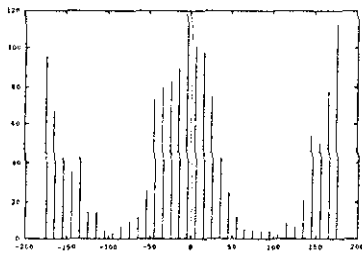


Figure 12:  $\psi$  of glycine

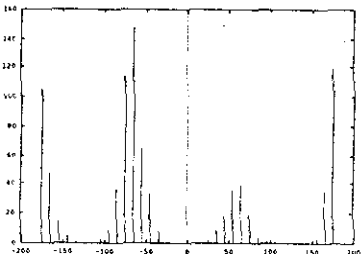


Figure 9:  $\chi_1$  of tyrosine

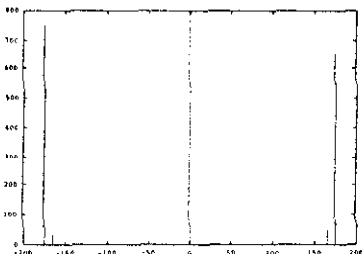
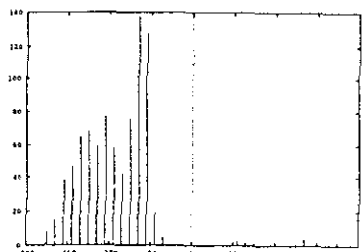
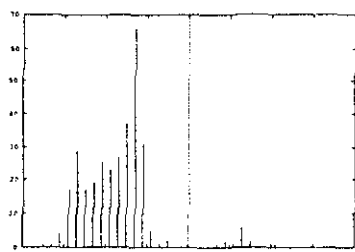
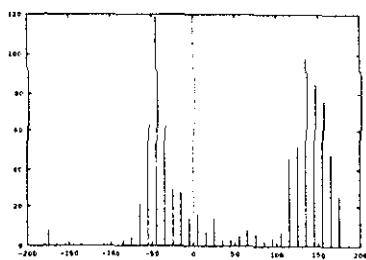
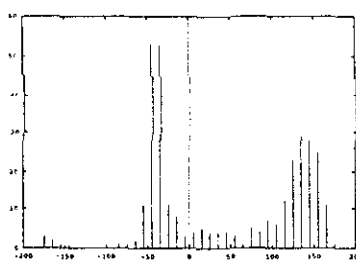
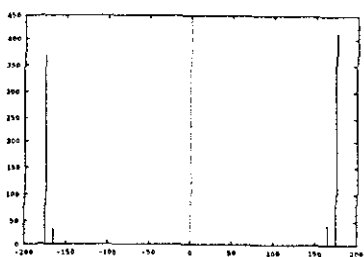
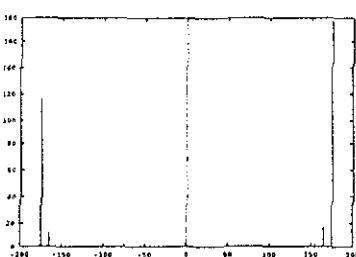
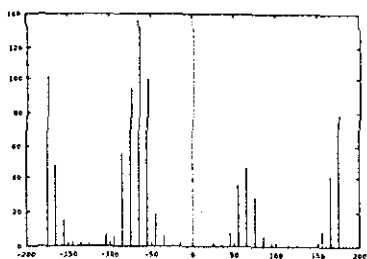
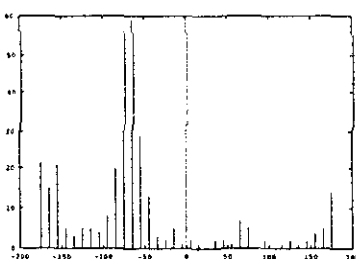


Figure 13:  $\omega$  of glycine

Figure 14:  $\phi$  of phenylalanineFigure 18:  $\phi$  of methionineFigure 15:  $\psi$  of phenylalanineFigure 19:  $\psi$  of methionineFigure 16:  $\omega$  of phenylalanineFigure 20:  $\omega$  of methionineFigure 17:  $\chi_1$  of phenylalanineFigure 21:  $\chi_1$  of methionine

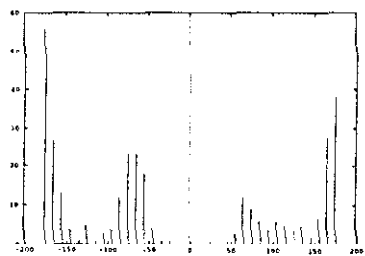


Figure 22:  $\chi_2$  of methionine

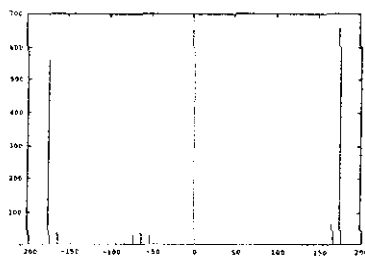


Figure 26:  $\omega$  of leucine

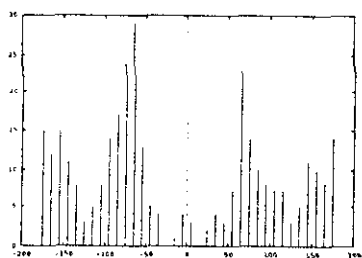


Figure 23:  $\chi_3$  of methionine

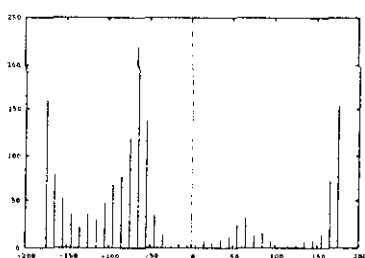


Figure 27:  $\chi_1$  of leucine

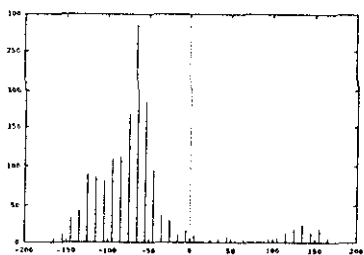


Figure 24:  $\phi$  of leucine

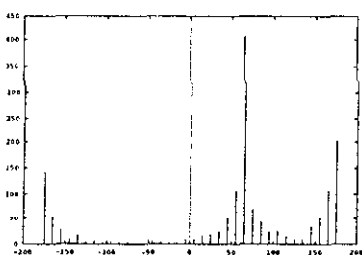


Figure 28:  $\chi_2$  of leucine

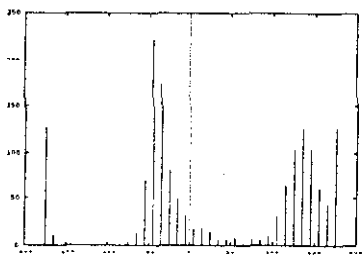


Figure 25:  $\psi$  of leucine

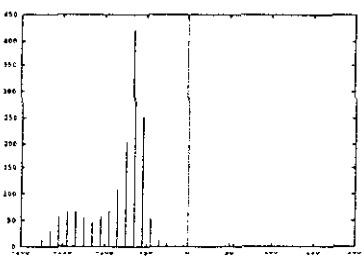
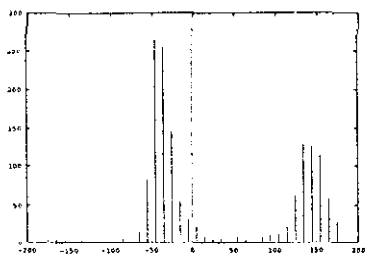
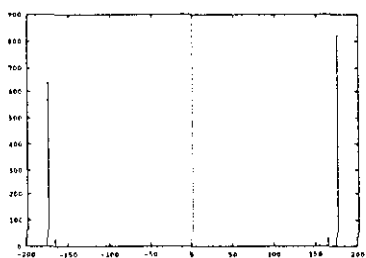
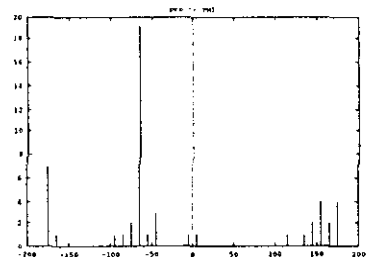
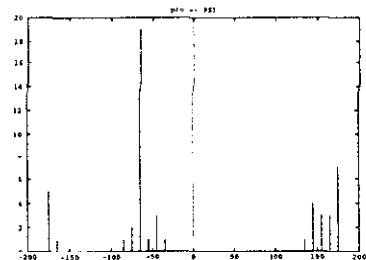
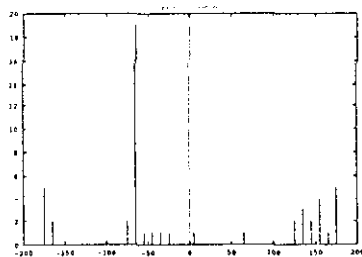


Figure 29:  $\phi$  of alanine

Figure 30:  $\psi$  of alanineFigure 31:  $\omega$  of alanineFigure 32:  $\phi$  of prolineFigure 33:  $\psi$  of prolineFigure 34:  $\omega$  of proline

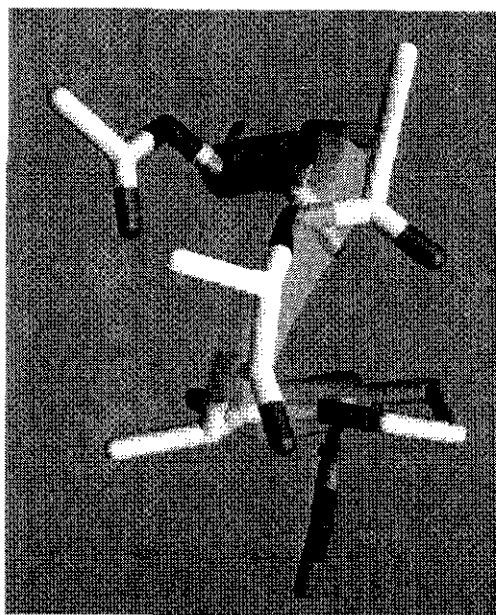


Figure 35: Plot of Ac-Ala<sub>4</sub>-P<sub>10</sub>-NHMe conformation,  $E^* = -18.910 \text{ kcal/mole}$

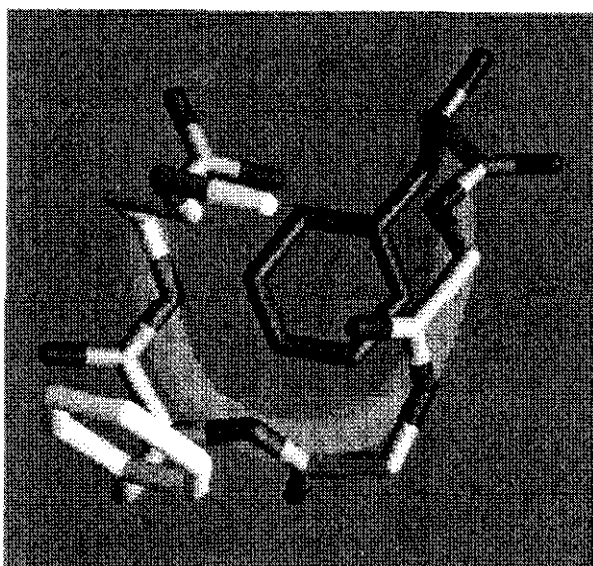


Figure 36: Plot of Met-enkephalin conformation,  $E^* = -11.707 \text{ kcal/mole}$

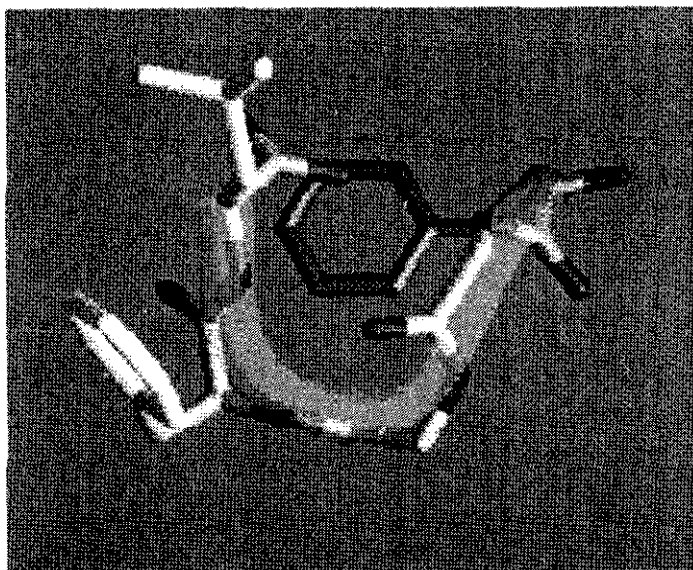


Figure 37: Plot of Leu-enkephalin conformation,  $E^* = -9.332 \text{ kcal/mole}$

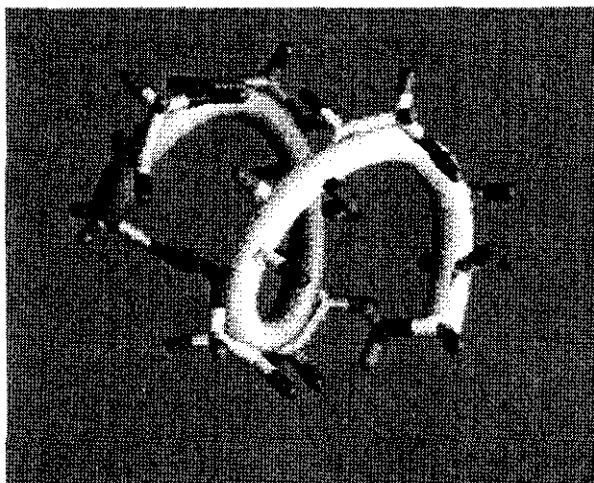


Figure 38: Plot of decaglycine conformation,  $E^* = -11.642 \text{ kcal/mole}$