Molecular Structure Prediction by Global Optimization

K.A. DILL
Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, CA 94118

A.T. PHILLIPS
Computer Science Department, United States Naval Academy, Annapolis, MD 21402

J.B. ROSEN
Computer Science and Engineering Department, University of California at San Diego, San Diego, CA 92093

Abstract. The CGU (convex global underestimator) global optimization method is used to predict the minimum energy structures, i.e. folded states, of small protein sequences. Computational results obtained from the CGU method applied to actual protein sequences using a detailed polypeptide model and a differentiable form of the Sun/Thomas/Dill potential energy function are presented. This potential function accounts for steric repulsion, hydrophobic attraction, and φ/ψ pair restrictions imposed by the so called Ramachandran maps. Furthermore, it is easily augmented to accommodate additional known data such as the existence of disulphide bridges and any other a priori distance data. The Ramachandran data is modeled by a continuous penalty term in the potential function, thereby permitting the use of continuous minimization techniques.

Keywords: Molecular conformation, protein folding, global optimization

1. Introduction

Macromolecules, such as proteins, require specific 3-dimensional conformations to function properly. These “native” conformations result primarily from intramolecular interactions between the atoms in the macromolecule, and also intermolecular interactions between the macromolecule and the surrounding solvent. Although the folding process appears to be quite complex, the instructions guiding this process are believed to be completely specified by the one-dimensional primary sequence of the protein or nucleic acid: external factors, such as helper (chaperone) proteins, present at the time of folding have no effect on the final state of the protein. Many denatured proteins and nucleic acids, for example, spontaneously refold into functional conformations once denaturing conditions are removed. Indeed, the existence of a *unique* native conformation, in which residues distant in sequence but close in proximity exhibit a densely packed hydrophobic core, suggests that this 3-dimensional structure is largely encoded within the

sequential arrangement of these hydrophobic (H) and polar (P) amino acids. The assumption that such hydrophobic interaction is the single most dominant force in the correct folding of a protein also suggests that simplified potential energy functions, for which the terms involve only pairwise H-H attraction and steric repulsion, may be sufficient to guide computational search strategies to the global minimum representing the native state.

Machine based prediction strategies, such as the one described in this paper, attempt to lessen the reliance on experts by developing a completely computational method. Such approaches are generally based on two assumptions. First, that there *exists* a potential energy function for the protein; and second that the folded state corresponds to the structure with the lowest potential energy (minimum of the potential energy function) and is thus in a state of thermodynamic equilibrium.

2. The Polypeptide Model

Computational search methods are not yet fast enough to find global optima in real-space representations using accurate all-atom models and potential functions. A practical conformational search strategy requires both a simplified, yet sufficiently realistic, molecular model with an associated potential energy function which consists of the dominant forces involved in protein folding, and also a global optimization method which takes full advantage of any special properties of this kind of energy function. In what follows, we describe such a model and an associated global optimization algorithm.

Each residue in the primary sequence of a protein is characterized by its backbone components NH-C_αH-C′O and one of 20 possible amino acid sidechains attached to the central C_α atom.The 3-dimensional structure of macromolecules is determined by internal molecular coordinates consisting of bond lengths l (defined by every pair of consecutive backbone atoms), bond angles θ (defined by every three consecutive backbone atoms), and the backbone dihedral angles φ, ψ, and ω, where φ gives the position of C′ relative to the previous three consecutive backbone atoms C′-N-C_α, ψ gives the position of N relative to the previous three consecutive backbone atoms N-C_α-C′, and ω gives the position of C_α relative to the previous three consecutive backbone atoms C_α-C′-N. Figure 2.1 illustrates this model.

Fortunately, these $9n$-6 parameters (for an n-residue structure) do not all vary independently. In fact, some of these ($7n$-4 of them) are regarded as fixed since they are found to vary within only a very small neighborhood of an experimentally determined value. Among these are the $3n$-1 backbone bond lengths l between the pairs of consecutive atoms N-C′ (fixed at 1.32 Å), C′-C_α (fixed at 1.53 Å), and C_α-N (fixed at 1.47 Å). Also, the $3n$-2 backbone bond angles θ defined by N-C_α-C′ (110˚), C_α-C′–N (114˚), and C′-N-C_α (123˚) are also fixed at

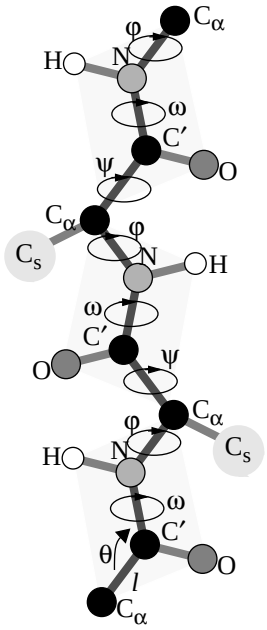

Figure 2.1 Polypeptide Model

their ideal values. Finally, the *n*-1 peptide bond dihedral angles ω are fixed in the trans (180˚) conformation. This leaves only the *n*-1 backbone dihedral angle pairs (φ,ψ) in the reduced representation model. These also are not completely independent; in fact, they are severely constrained by known chemical data (the Ramachandran plot) for each of the 20 amino acid residues.

Furthermore, since the atoms from one C_α to the next C_α along the backbone can be grouped into rigid *planar* peptide units, there are no extra parameters required to express the 3-dimensional position of the attached O and H peptide atoms. These bond lengths and bond angles are also known and fixed at 1.24 Å and 121˚ for O, and 1.0 Å and 123˚ for H.

A key element of this simplified polypeptide model is that each sidechain is classified as either hydrophobic or polar, and is represented by only a single "virtual" center of mass atom. Since each sidechain is represented by only the single center of mass "virtual atom" C_s, no extra parameters are needed to define the position of each sidechain with respect to the backbone mainchain. The twenty amino acids are thus classified into two groups, hydrophobic and polar, according to the scale given by Miyazawa and Jernigan in [4].

Corresponding to this simplified polypeptide model is a potential energy function also characterized by its simplicity. This function includes just three components: a contact energy term favoring pairwise H-H residues, a steric repulsive

term which rejects any conformation that would permit unreasonably small interatomic distances, and a main chain torsional term that allows only certain preset values for the backbone dihedral angle pairs (φ,ψ). Since the residues in this model come in only two forms, H (hydrophobic) and P (polar), where the H-type monomers exhibit a strong pairwise attraction, the lowest free energy state is obtained by those conformations with the greatest number of H-H "contacts" (see [1], [7]). Despite its simplicity, the use of this type of potential function has already proven successful in studies conducted independently by Sun, Thomas, and Dill [8] and by Srinivasan and Rose [6]. Both groups have demonstrated that this type of potential function is sufficient to accurately model the forces which are most responsible for folding proteins. The specific potential function used initially in this study is a simple modification of the Sun/Thomas/Dill energy function and has the following form:

$$E_{total} = E_{ex} + E_{hp} + E_{\varphi\psi} \tag{1}$$

where E_{ex} is the steric repulsive term which rejects any conformation that would permit unreasonably small interatomic distances, E_{hp} is the contact energy term favoring pairwise H-H residues, and $E_{\varphi\psi}$ is the main chain torsional term that allows only those (φ,ψ) pairs which are permitted by the Ramachandran plot. In particular, the excluded volume energy term E_{ex} and the hydrophobic interaction energy term E_{hp} are defined in this case as follows:

$$E_{ex} = \sum_{ij} \frac{C_1}{1.0 + exp((d_{ij} - d_{eff})/d_w)}, \text{ and}$$

$$E_{hp} = \sum_{|i-j|>2} \varepsilon_{ij} f(d_{ij}) \text{ where } f(d_{ij}) = \frac{C_2}{1.0 + exp((d_{ij} - d_0)/d_t)}.$$

The excluded volume term E_{ex} is a soft sigmoidal potential where d_{ij} is the interatomic distance between two C_α atoms or between two sidechain center of mass atoms C_s, d_w = 0.1 Åwhich determines the rate of decrease of E_{ex}, d_{eff} = 3.6 Å for C_α atoms and 3.2 Å for the sidechain centroids which determine the midpoint of the function (i.e. where the function equals 1/2 of its maximum value). The constant multiplier C_1 was set to 5.0 which determines the hardness of the sphere in the excluded volume interaction. Similarly, the hydrophobic interaction energy term E_{hp} is a short ranged soft sigmoidal potential where d_{ij} represents the interatomic distance between two sidechain centroids C_s, d_0 = 6.5 Å and d_t = 2.5 Å which represent the rate of decrease and the midpoint of E_{hp}, respectively. The hydrophobic interaction coefficient ε_{ij} = -1.0 when both residues *i* and *j* are hydrophobic, and is set to 0 otherwise. The constant multiplier C_2 = 1.0 determines the interaction value and is the equivalent of 1/5 of one excluded volume violation. The model is not very sensitive to the pair of constants C_1 and C_2 provided that C_1

is larger than C_2. Figure 2.2 shows the combined effect of the energy terms E_{ex} +

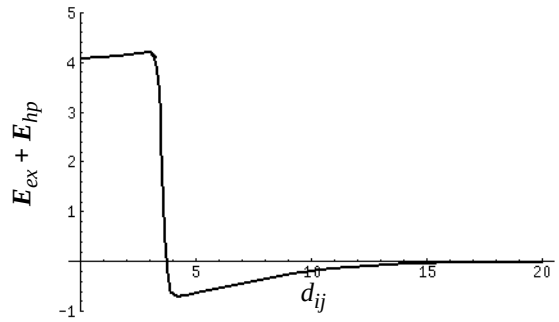

Figure 2.2 Combined Potential Function Energy Terms $E_{ex} + E_{hp}$

E_{hp} for a pair of H-H residues.

The final term in the potential energy function, $E_{\varphi\psi}$, is the torsional penalty term allowing only "realistic" (φ,ψ) pairs in each conformation. That is, since φ and ψ refer to rotations of two rigid peptide units around the same C_α atom (see Figure 2.1), most combinations produce steric collisions either between atoms in different peptide groups or between a peptide unit and the side chain attached to C_α (except for glycine). Hence, only certain specific combinations of (φ,ψ) pairs are actually observed in practice, and are often conveyed via the Ramachandran plot, such as the one in Figure 2.3, and the φ-ψ search space is therefore very

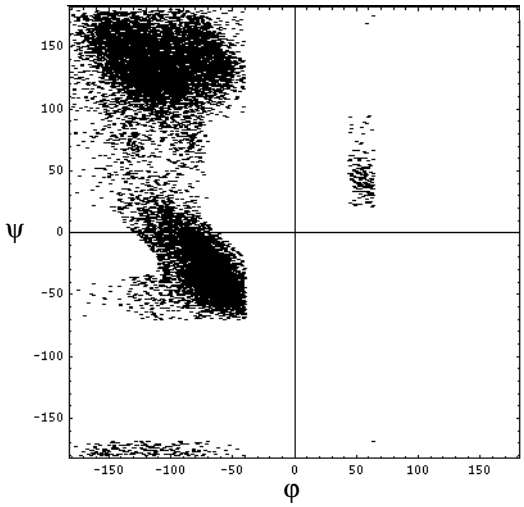

Figure 2.3 Ramachandran Plot for All Residues Except GLY and PRO

much restricted.

Use of the Ramachandran plot in determining protein structure is essential in order to restrict the φ-ψ plane to those regions observed for actual protein molecules. While most other global optimization methods for protein structure prediction also depend on this property, they are invariably forced to use some variant of simulated annealing (or any other method based on random sampling rather than gradient information) to perform the optimization due to the lack of a smooth, i.e. differentiable, representation for $E_{\varphi\psi}$. This, in turn, results in a very slow local minimization process. Our approach, however, is to model the Ramachandran data by a smooth function which will have the approximate value zero in any permitted region, and a large positive value in all excluded regions. This "penalty term" is therefore differentiable and will be easily computed.

A key observation in the construction of the function $E_{\varphi\psi}$ is that the set of allowable (φ,ψ) pairs form compact clusters in the φ-ψ plane. By enclosing each such cluster in an appropriately constructed ellipsoid, we may use the ellipsoids to define the energy term $E_{\varphi\psi}$. In particular, given p regions (ellipsoids) R_1, R_2,..., R_p, containing the experimentally allowable (φ,ψ) pairs (see Figure 2.4), we desire

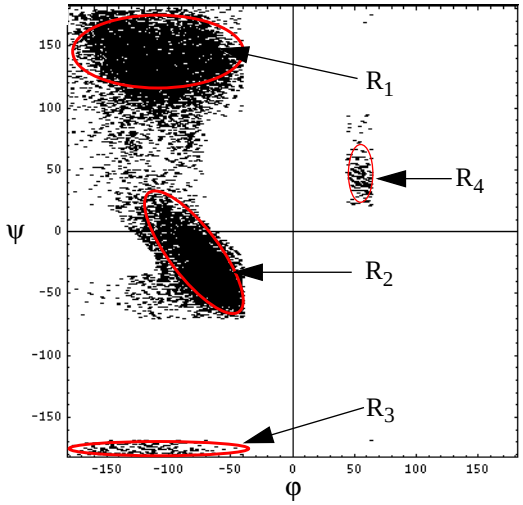

Figure 2.4 Approximating Ramachandran Data by Ellipsoids

the energy term $E_{\varphi\psi}$ to satisfy

$$E_{\varphi\psi} \cong \begin{cases} 0 \text{ if } (\varphi, \psi) \in R_i \text{ for some } i \\ \beta \qquad \text{otherwise} \end{cases} \tag{2}$$

where β is some large constant penalty. To obtain such an energy term, we first represent the i^{th} ellipsoid R_i by a quadratic function $q_i(\varphi,\psi)$ which is positive definite (both eigenvalues positive) and satisfies $q_i(\varphi,\psi) = 0$ on the boundary of the

ellipsoid R_i, $q_i(\varphi,\psi) < 0$ in the interior of R_i, and $q_i(\varphi,\psi) > 0$ in the exterior. By simply constructing a sigmoidal penalty term of the form

$$E_{\varphi\psi} = \frac{\beta}{1 + \sum_{i=1}^{p} exp(-\gamma_i q_i(\varphi, \psi))} \tag{3}$$

where the constants $\gamma_i > 0$ determine the rate by which $E_{\varphi\psi}$ approaches 0 or β near an ellipsoid boundary, then it is easy to see that $E_{\varphi\psi} \cong 0$ in the ellipsoid's interior, and $E_{\varphi\psi} \cong \beta$ at distant exterior points, thus satisfying Eq (2). Figure 2.5shows $E_{\varphi\psi}$

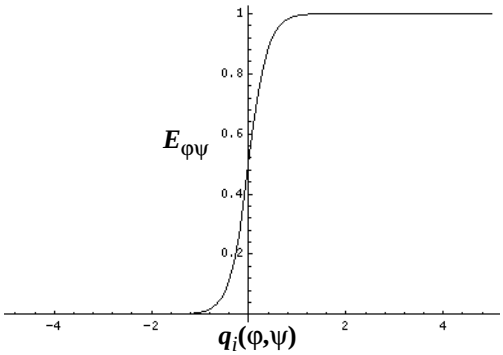

Figure 2.5 Sigmoidal Penalty Term $E_{\varphi\psi}$ for β = 1 and $\gamma_1 = 5$

as a function of $q_i(\varphi,\psi)$ for a single ellipsoid with the values of β = 1 and $\gamma_1 = 5$. Figure 2.6 illustrates the 3-dimensional plot of $E_{\varphi\psi}$ (for β = 1, and all $\gamma_i = 100$) for the data provided in Figure 2.4. Likewise, Figure 2.7 shows the 2-dimensional topographical map for that same function.

Figures 2.8 and 2.9 show the Ramachandran plots for GLY and PRO, respectively. The corresponding 3-dimensional plots of $E_{\varphi\psi}$ for these same residues are shown in Figures 2.10 and 2.11. Since these two residues differ substantially from the other 18 residues in their Ramachandran data, they require different forms for the penalty term $E_{\varphi\psi}$. In summary, although there are 20 different amino acid residues, since 18 of these exhibit very similar torsional, i.e. (φ,ψ), distributions, only three forms of the torsional penalty term $E_{\varphi\psi}$ are required: one for GLY, one for PRO, and one for the remaining 18 residues.

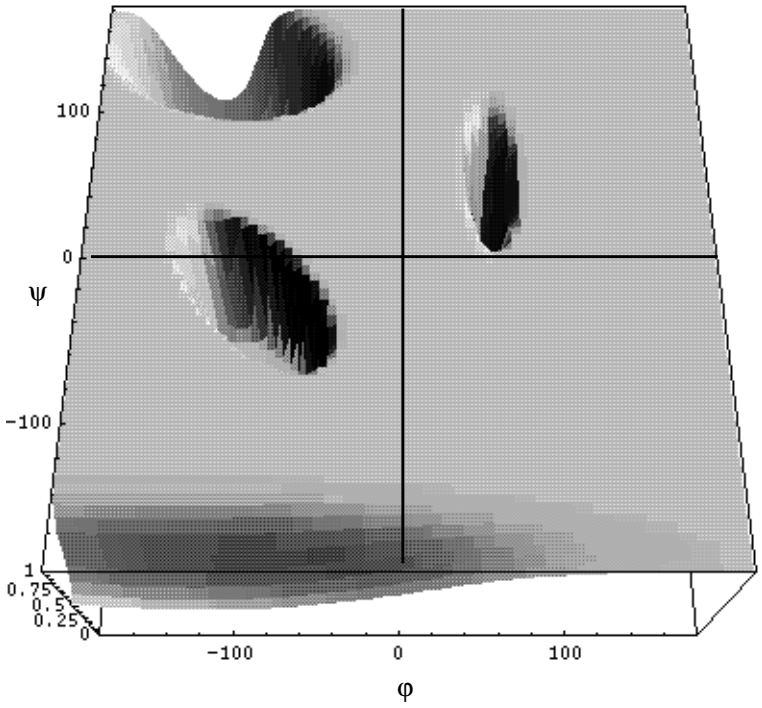

Figure 2.6 3-Dimensional Plot of $E_{\varphi\psi}$ (for $\beta = 1.0$) Corresponding to the Ramachandran Data in Figure 2.4

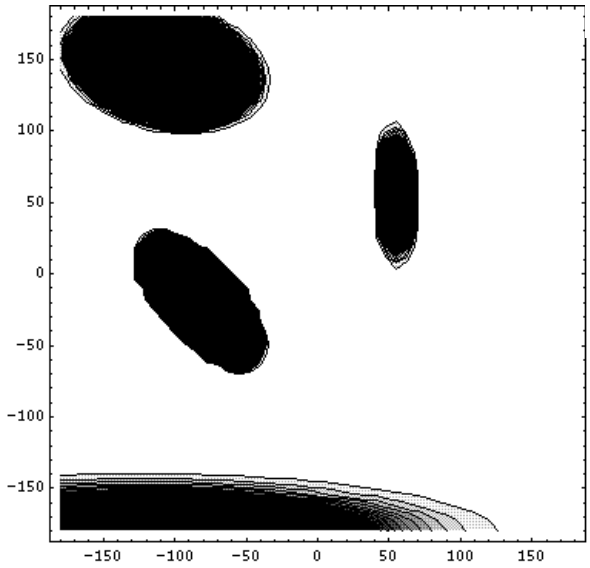

Figure 2.7 2-Dimensional Topographical Map Corresponding to the Function Shown in Figure 2.6

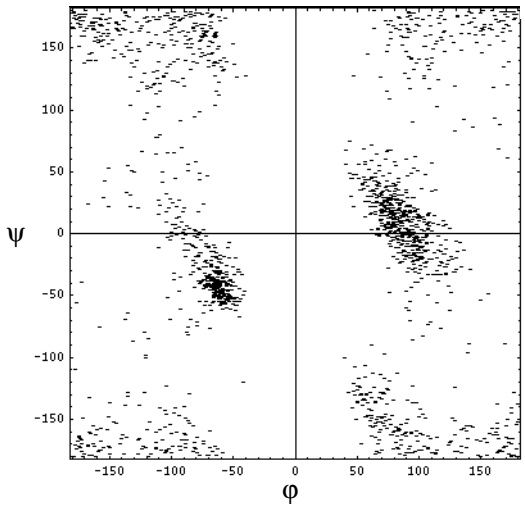

Figure 2.8 Ramachandran Plot for the Residue GLY

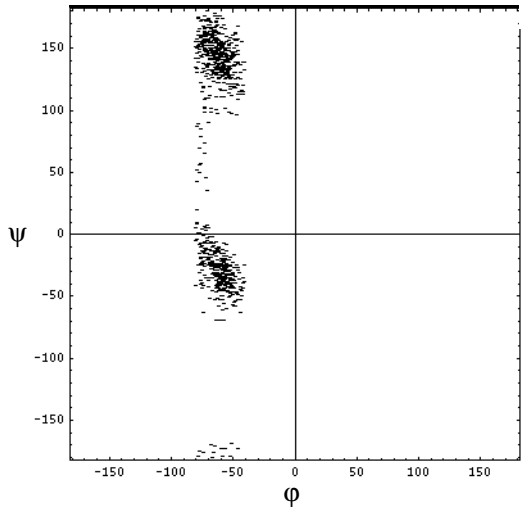

Figure 2.9 Ramachandran Plot for the Residue PRO

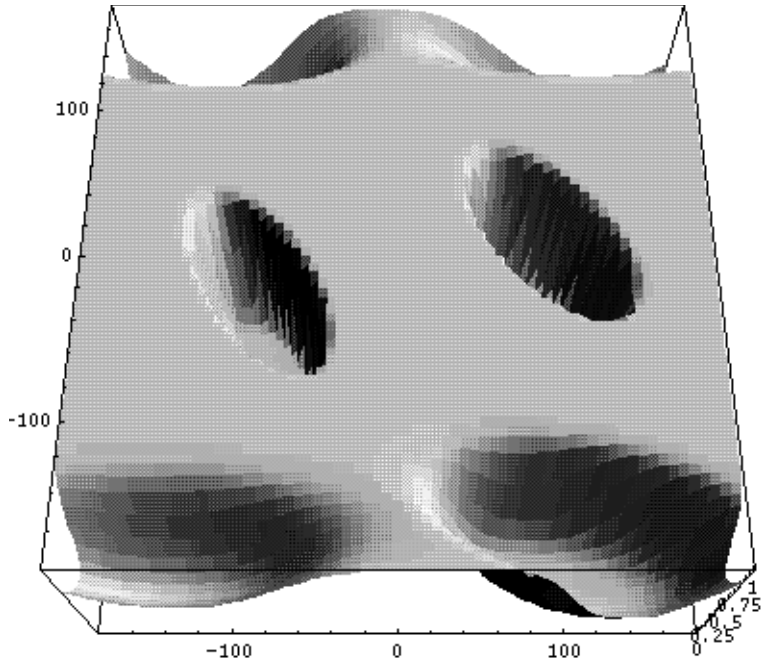

Figure 2.10 3-Dimensional Plot of $E_{\varphi\psi}$ (β = 1.0) Corresponding to the Data for GLY in Figure 2.8

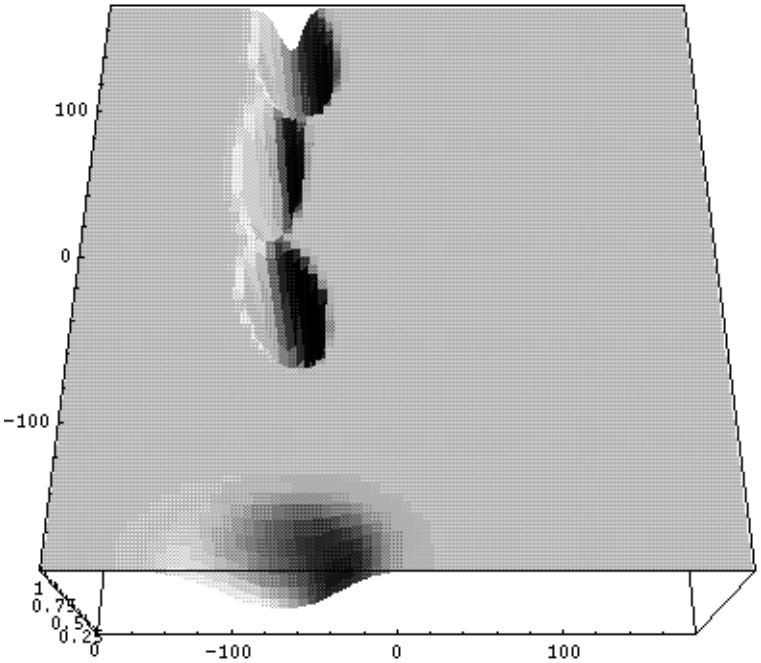

Figure 2.11 3-Dimensional Plot of $E_{\varphi\psi}$ (β = 1.0) Corresponding to the Data for PRO in Figure 2.9

3. The CGU Global Optimization Algorithm

One practical means for finding the global minimum of the polypeptide's potential energy function is to use a global underestimator to localize the search in the region of the global minimum. This CGU (convex global underestimator) method is designed to fit all known local minima with a convex function which underestimates all of them, but which differs from them by the minimum possible amount in the discrete L_1 norm (see Figure 3.1). The minimum of this underestimator is

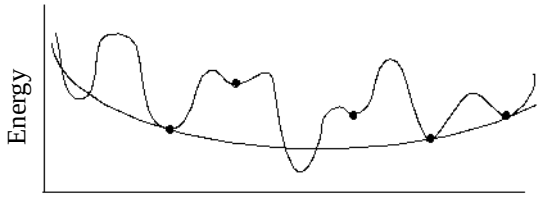

Figure 3.1 The Convex Global Underestimator (CGU)

used to predict the global minimum for the function, allowing a more localized conformer search to be performed based on the predicted minimum (see Figures 3.2 and 3.3). A new set of conformers generated by the localized search then

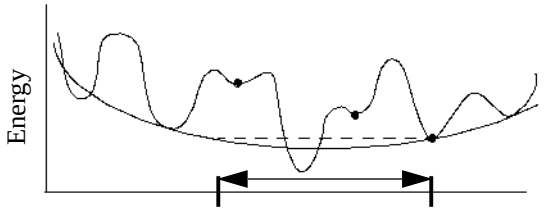

Figure 3.2 Defining the New Search Domain

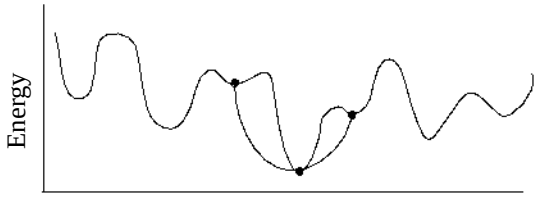

Figure 3.3 The New CGU Over the Reduced Search Domain

serves as a basis for another quadratic underestimation over the reduced space. After several repetitions, the global minimum can be found with reasonable assurance.

This CGU method, first described in [5], has previously been applied successfully to a simpler "string of beads" model as described in [2]. In this paper, we

apply the CGU algorithm to the more realistic polypeptide model shown in Figure 2.1.

4. Results for Small Polypeptides

The CGU algorithm was tested on five small polypeptides: met-enkephalin, bradykinin, oxytocin, mellitin, and PSU-SEQ-9. To assess the accuracy of the CGU computed structures as compared to the known structures, the distance matrix error

$$\text{DME} = \sqrt{\left(\frac{2}{N(N-1)}\right)\sum_{i=1}^{N}\sum_{j=i+1}^{N}(r_{ij}-r_{ij}^{c})^{2}}$$

was computed, where the pairwise distances r_{ij} are calculated over all C_α backbone atoms, and the superscript "c" indicates the "correct" target conformation (usually obtained from the Brookhaven Protein Database).

Met-enkephalin, first used in computational protein folding studies by Scheraga's group [3], is a small brain peptide that is a natural ligand for opiate receptor cites. It is often used as an initial test case for folding algorithms because it is small yet non-trivial. Met-enkephalin consists of only five residues (TYR-GLY-GLY-PHE-MET), of which only three (TYR, PHE, and MET) are hydrophobic. For this test case, the global minimum energy obtained by the CGU algorithm in 73 seconds (wall clock time) on an eight processor Dec Alpha workstation cluster was -43.79 kcal/mol (using the modified Sun/Thomas/Dill energy function previously described). Figures 4.1 and 4.2 show the CGU computed "ball-and-stick"

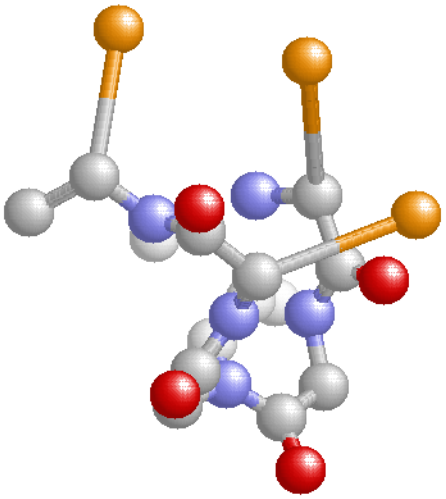

Figure 4.1 The CGU Determined Native Structure for Met-Enkephalin: Ball-and-Stick Representation

Figure 4.2 The CGU Determined Native Structure f
Met-Enkephalin: Ribbon Representation

and “ribbon” representations for this peptide (as seen from the same viewpoint), respectively.

Bradykinin, which is a hormone-like peptide that inhibits inflammatory reactions, consists of nine residues (ARG-PRO-PRO-GLY-PHE-SER-PRO-PHE-ARG), of which only two are hydrophobic (both PHE residues). The global minimum energy obtained by the CGU algorithm in 382 seconds (6.4 minutes wall clock time) on the eight processor Dec Alpha workstation cluster was -21.89 kcal/mol. Figures 4.3 and 4.4 show the CGU computed representations for this peptide.

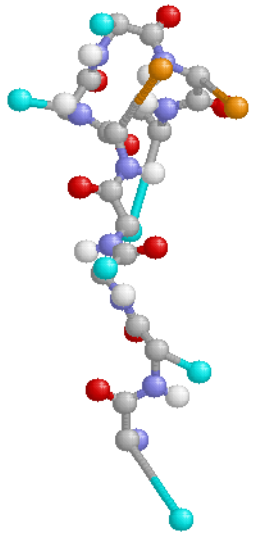

Figure 4.3 The CGU Determined Native Structure for
Bradykinin: Ball-and-Stick Representation

Oxytocin is a pituitary hormone consisting of nine residues (CYS-TYR-ILE-GLN-ASN-CYS-PRO-LEU-GLY) with a disulphide bridge between the CYS residues (1 and 6). Five of the nine residues are hydrophobic (CYS, TYR, ILE, CYS again, and LEU). The existence of the disulphide bridge greatly reduces the conformation space since the distance between the two CYS residues is effectively

Figure 4.4 The CGU Determined Native Structure for Bradykinin: Ribbon Representation

fixed (and the CGU algorithm makes use of this property, see [2]). The global minimum energy obtained by the CGU algorithm in 445 seconds (7.4 minutes wall clock time) on the eight processor Dec Alpha workstation cluster was -119.01 kcal/mol. The DME error for this structure (compared with "pdb1xy1.ent" from the Brookhaven Protein Database) was 0.338 Ang. Figures 4.5 and 4.6 show

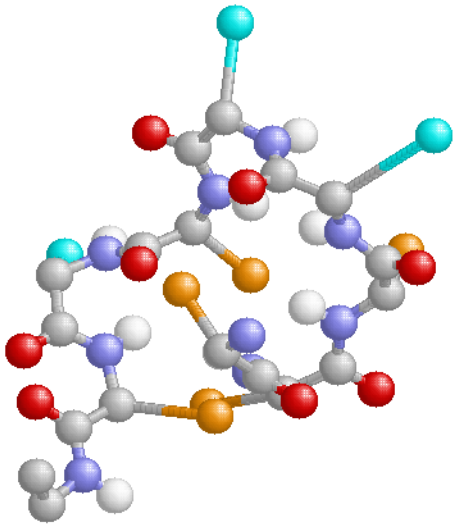

Figure 4.5 The CGU Determined Native Structure for Oxytocin: Ball-and-Stick Representation

the CGU computed representations for this peptide.

Mellitin consists of 27 residues (GLY-ILE-GLY-ALA-VAL-LEU-LYS-VAL-LEU-THR-THR-GLY-LEU-PRO-ALA-LEU-ILE-SER-TRP-ILE-LYS-ARG-LYS-ARG-GLN-GLN-GLY) of which twelve are hydrophobic (ILE, ALA, VAL,

Figure 4.6 The CGU Determined Native Structure for Oxytocin: Ribbon Representation

LEU, and TRP). It is commonly found in bee venom and is responsible for an increase in cell permeability. For this peptide, the global minimum energy obtained by the CGU algorithm in 49692 seconds (13.8 hours wall clock time) on the eight processor Dec Alpha workstation cluster was -903.38 kcal/mol. The DME error for this structure (compared with “pdb2mlt.ent” from the Brookhaven Protein Database) was 0.394 Ang. Figures 4.7 and 4.8 show the CGU computed

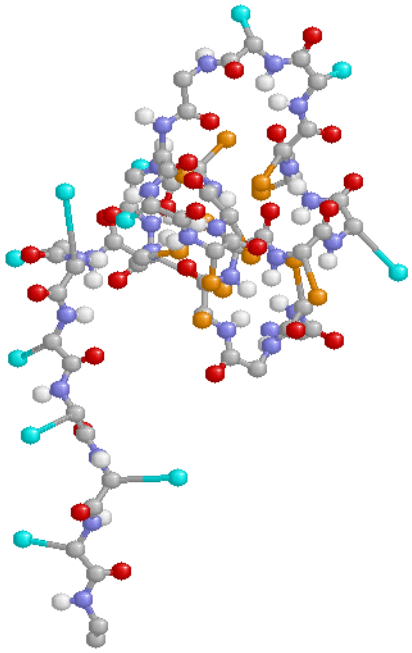

Figure 4.7 The CGU Determined Native Structure for Mellitin: Ball-and-Stick Representation

representations for this peptide.

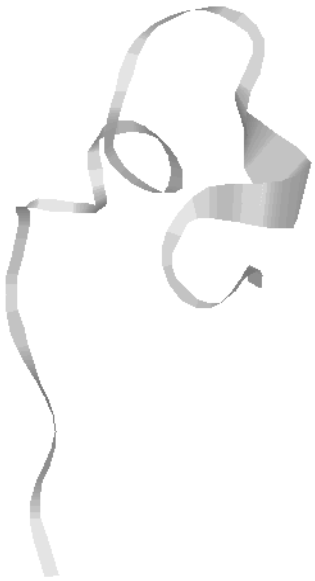

Figure 4.8 The CGU Determined Native Structure for Mellitin: Ribbon Representation

PSU-SEQ-9 is not a naturally occurring peptide, but rather is a subpart of a larger structure used in binding sperm to egg membranes. In this case, the native structure for PSU-SEQ-9 is not known. The sixteen residue sequence (LEU-TYR-PRO-GLN-ASP-ARG-PRO-ARG-SER-GLN-PRO-GLN-PRO-LYS-ALA-ASN) for PSU-SEQ-9 involves only three hydrophobic residues (LEU, TYR, and ALA), and the global minimum energy obtained by the CGU algorithm in 4488 seconds (1.25 hours wall clock time) on the eight processor Dec Alpha workstation cluster was -43.78 kcal/mol. Figures 4.9 and 4.10 show the CGU computed representa-

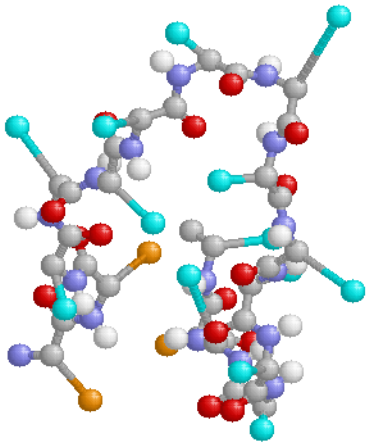

Figure 4.9 The CGU Determined Native Structure for PSU-SEQ-9: Ball-and-Stick Representation

Figure 4.10 The CGU Determined Native Structure for PSU-SEQ-9: Ribbon Representation

tions for this peptide.

5. Conclusions

The CGU method has been shown to be both practical and effective in computing the minimum energy structures for the small protein sequences tested. In those cases for which a global solution is known, the DME between the CGU computed and known structures is always observed to be less than 0.5 Ang. Furthermore, the use of a differentiable representation of $E_{\varphi\psi}$ is crucial to permit the CGU method (and, in fact, any other method based on continuous minimization) to proceed. This penalty function approach to representing the Ramachandran data is a key component of the modified Sun/Thomas/Dill potential function. Improvements in the CGU algorithm, which are currently being investigated, should reduce the computation times substantially. This will permit the application of the techniques described here to the study of larger protein molecules.

Acknowledgments

K.A. Dill was supported by the NSF grant BIR-9119575, A.T. Phillips was supported by NIH grant RR-08605 and the San Diego Supercomputer Center, and J.B. Rosen was supported by the ARPA/NIST grant 60NANB2d1272 and NSF grant CCR-9509085

References

1. K.A. Dill, *Dominant Forces in Protein Folding*, Biochemistry **29** (1990), 7133-7155.
2. K.A. Dill, A.T. Phillips, and J.B. Rosen, *CGU: An Algorithm for Molecular Structure Prediction*, IMA Volumes in Mathematics and its Applications, in press (1996).
3. Z. Li, and H.A. Scheraga, *Monte-Carlo Minimization Approach to the Multiple-Min-*

ima Problem in Protein Folding, Proceedings of the National Academy of Sciences USA **84** (1987): 6611-6615.

4. S. Miyazawa, and R.L. Jernigan, *A New Substitution Matrix for Protein Sequence Searches Based on Contact Frequencies in Protein Structures*, Protein Engineering **6** (1993): 267-278.
5. A.T. Phillips, J.B. Rosen, and V.H. Walke, *Molecular Structure Determination by Convex Global Underestimation of Local Energy Minima*, Dimacs Series in Discrete Mathematics and Theoretical Computer Science **23** (1995), P.M. Pardalos, G.-L. Xue, and D. Shalloway (Eds), 181-198.
6. R. Srinivasan and G.D. Rose, *LINUS: A Hierarchic Procedure to Predict the Fold of a Protein*, PROTEINS: Structure, Function, and Genetics **22** (1995), 81-99.
7. S. Sun, *Reduced representation model of protein structure prediction: statistical potential and genetic algorithms*, Protein Science **2** (1993), 762-785.
8. S. Sun, P.D. Thomas, and K.A. Dill, *A Simple Protein Folding Algorithm using a Binary Code and Secondary Structure Constraints*, Protein Engineering, submitted (1995).