Protein Structure and Energy Landscape Dependence on Sequence using a Continuous Energy Function

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. We have recently described a new conformational search strategy for protein folding algorithms, called the CGU (convex global underestimator) method. Here we use a simplified protein chain representation and a differentiable form of the Sun/Thomas/Dill energy function to test the CGU method. Standard search methods, such as Monte Carlo and molecular dynamics are slowed by kinetic traps. That is, the computer time depends more strongly on the shape of the energy landscape (dictated by the amino acid sequence) than on the number of degrees of freedom (dictated by the chain length). The CGU method is not subject to this limitation, since it explores the underside of the energy landscape, not the top. We find that the CGU computer time is largely independent of the monomer sequence, for different chain folds, and scales as $O(n^4)$ with chain length. By using different starting points, we show that the method appears to find global minima. Since we can currently find stable states of 36-residue chains in 2.4 hours, the method may be practical for small proteins.

Keywords: Molecular conformation, protein folding, global optimization

1. Introduction

To develop a computer algorithm that will predict the native structure of a protein given only its amino acid sequence requires three ingredients: a suitable chain representation, an accurate energy function, and a fast search strategy that can find the globally optimal lowest energy conformations. Our aim here is not to attempt to solve all these problems or to develop a protein folding algorithm. Our aim is more modest. Here we use an imperfect simplified chain representation and energy function to test a new conformational search method. Our model is a differentiable form of the Sun/Thomas/Dill chain representation and energy function. It is protein-like in the following respects: it includes chain connectivity,

steric constraints, φψ preferences, and hydrophobic and hydrogen bonding interactions. Hence chains collapse to compact states with hydrophobic cores and some hydrogen bonded structure. The model, however, is not sufficient to predict native folds correctly from amino acid sequences, even though it does well in some limited tests [10]. Thus our aim is not to fold proteins; our aim is to test a method of searching for lowest energy states in a protein-like model.

Standard search methods, including molecular dynamics, Monte Carlo, simulated annealling, and genetic algorithms, often get stuck in kinetic traps. The challenge in developing fast search strategies is in understanding what slows them down. Proteins are difficult for two reasons. First, we require the global optimum, rather than one of the large number of local optima. Second, an algorithm must be able to deal with very different shapes of energy landscapes, and the shape is often a stronger determinant of computer time than the size of the landscape. The shape of the landscape is dependent on the amino acid sequence. The size is dependent on the chain length. Often there is no simple scaling with the chain length.

Our current method takes a rather different approach. Rather than meandering over the tops of energy landscapes, our CGU method burrows under it, and should therefore not be affected by the heights of kinetic barriers. As a test, here we explore the dependence of the CGU search time on chain length and monomer sequence. Also we vary the starting conformations in order to check that the method is finding global optima, rather than local optima. The method does not get stuck in kinetic traps, and appears to find global optima in a time that depends on n^4, where n is the chain length, independently of the monomer sequence. It may be a practical search method for more realistic energy functions.

2. Protein Model and Global Search Strategy

Practical search strategies for the protein folding problem currently require a simplified, yet sufficiently realistic, molecular model with an associated potential energy function representing the dominant forces involved in protein folding [9]. In our present model, 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 three-dimensional structure of the chain is determined by internal molecular coordinates consisting of bond lengths *l*, bond angles θ, and the backbone dihedral angles φ, ψ, and ω. Fortunately, these 9*r*-6 parameters (for an *r*-residue structure) do not all vary independently. Some of these (7*r*-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 3*r*-1 backbone bond lengths *l*, the 3*r*-2 backbone bond angles θ, and the *r*-1 peptide bond dihedral angles ω (fixed in the trans conformation). This leaves only the *r*-1 backbone dihedral angle pairs (φ,ψ) in the reduced representation model. These also are not completely independent; 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 three-dimensional position of the attached O and H peptide atoms. Hence, these bond lengths and bond angles are also known and fixed. Figure 1 illustrates this model.

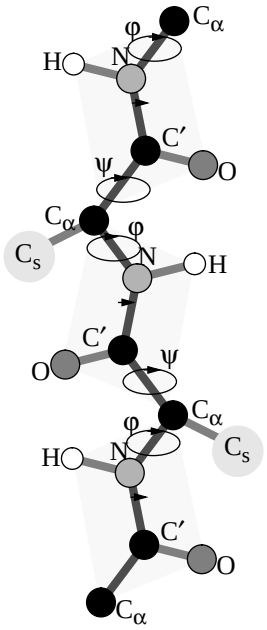

Figure 1 Simple Polypeptide Model

Another 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 [5].

Corresponding to this simplified polypeptide model is a simple energy function. This function includes four components: a contact energy term favoring pairwise hydrophobic residues, a second contact term favoring hydrogen bond formation between donor NH and acceptor C′=O pairs, 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, hydrophobic and polar, where the hydrophobic monomers

exhibit a strong pairwise attraction, the lowest free energy state involves those conformations with the greatest number of hydrophobic "contacts" [1] and intrastrand hydrogen bonds. Simplified potential function have been successful in studies by Sun, Thomas, and Dill [10], by Srinivasan and Rose [7], and by Yue and Dill [11]. We used here a simple modification of the Sun/Thomas/Dill energy function.

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 2). The minimum of this underestimator is used to predict the global minimum for the function, allowing a more localized conformer search to be performed based on the predicted minimum. More precisely, given an *r*-residue structure with n=2r-2 backbone dihedral angles, we denote a conformation of this simplified model by $\phi \in \mathbf{R}^n$, and the corresponding simplified potential energy function value by $F(\phi)$. Then, assuming that $k \geq 2n+1$ local minimum conformations $\phi^{(j)}$, for j=1,...,k, have been computed, a convex quadratic underestimating function $\Psi(\phi)$ is fitted to these local minima so that it underestimates all the local minima, and normally interpolates $F(\phi^{(j)})$ at 2n+1 points. This is accomplished by determining the coefficients in the function $\Psi(\phi)$ so that

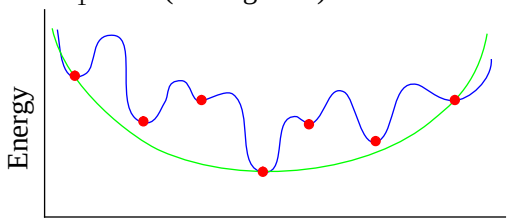

Figure 2 The Convex Global Underestimator (CGU)

$$\delta_j = F(\phi^{(j)}) - \Psi(\phi^{(j)}) \geq 0 \tag{1}$$

for j=1,...,k, and where $\sum_{j=1}^{k} \delta_j$ is minimized. That is, the difference between $F(\phi)$ and $\Psi(\phi)$ is minimized in the discrete L_1 norm over the set of *k* local minima $\phi^{(j)}$, j=1,...,k. The underestimating function $\Psi(\phi)$ used in this CGU method is given by

$$\Psi(\phi) = c_0 + \sum_{i=1}^{n} \left(c_i \phi_i + \frac{1}{2} d_i \phi_i^2 \right). \tag{2}$$

Note that c_i and d_i appear linearly in the constraints of (1) for each local minimum $\phi^{(j)}$. Convexity of this quadratic function is guaranteed by requiring that $d_i \geq$

0 for i=1,...,n. Other linear combinations of convex functions could also be used, but this quadratic function is the simplest.

A new set of conformers generated by the localized search then serves as a basis for another quadratic underestimation over the reduced space. After several repetitions, the global minimum conformation ϕ_G and its associated global minimum energy $F(\phi_G)$ can be found with reasonable assurance. For more specific details of the CGU method and its computational results, see [2], [3], [4], and [6].

3. Global Underestimation of the Energy Landscape

As summarized in the previous section, the CGU algorithm will determine a global minimum backbone torsion angle vector ϕ_G and corresponding global minimum energy function value $F_G = F(\phi_G)$. As part of the CGU algorithm, a relatively large number of local minima $\phi^{(j)}$, j =1,...,k, of the function $F(\phi)$ will also be computed. We denote the corresponding function values by $F_j = F(\phi^{(j)}) \geq F_G$ for j=1,...,k. Typically we have k ≥ 2n+1, where n is the number of backbone torsion angles since this is a necessary condition for constructing the minimum L1 convex global underestimator. Using all of these local minima, a final convex quadratic global underestimating function is determined, similar to (2), by solving a linear program formulated so that ϕ_G, the global minimum of the potential function $F(\phi)$, is also the global minimum of the new global underestimating $\Psi(\phi)$, and so $\Psi(\phi_G) = F_G$. The coefficients d_i of this final underestimating function are determined by constructing the underestimating function such that

$$\Psi(\phi) = F_G + \frac{1}{2}(\phi - \phi_G)^T D(\phi - \phi_G) \tag{3}$$

$$= F_G + \frac{1}{2}\sum_{i=1}^{n} d_i(\phi_i - \phi_{Gi})^2$$

where the diagonal matrix $D=diag(d_i) \in \mathbf{R}^{nxn}$, and then solving the linear program

$$\min_{d_i} \sum_{j=1}^{k} \delta_j \tag{4}$$

subject to

$$\delta_i = F_j - \Psi(\phi^{(j)}) \geq 0 \text{ and } 0 \leq d_i \leq d_{max} \text{ for i=1,...,n.}$$

The value d_{max} is a large specified upper bound. This prevents the underestimating function from increasing too rapidly as a function of the deviation of any torsion angle ϕ_i from its global minimum value $(\phi_G)_i$.

The solution to (4) will have the property that at least 2n+1 of the local minima F_j will be interpolated by the underestimator $\Psi(\phi)$. All remaining local minima

will be strictly underestimated, but will differ from $\Psi(\phi)$ by the minimum possible, as measured in the L1 norm. A coefficient d_i can be zero only if the global minimum function value F_G is attained at two different conformations ϕ. The CGU algorithm eliminates isomers, so except for very small molecules, we do not observe this. Therefore, typically all $d_i > 0$.

On the other hand, if a particular torsion angle, say ϕ_i, has the value $(\phi_G)_i$ for *every* local minimum, then the corresponding $d_i = d_{max}$. This simply means that there is a large penalty for changing ϕ_i from its global minimum value $(\phi_G)_i$.

The true energy landscape can be thought of as a surface above an n-dimensional horizontal hyperplane, with each point in the hyperplane representing a conformation ϕ. The energy $F(\phi)$ - F_G is then represented by the height of the surface above the hyperplane, as given by the (n+1) coordinate. Each local minimum $\phi^{(j)}$ of $F(\phi)$ has an (n+1) coordinate value of F_j- F_G.

It is important to note that computing many local minima is a crucial aspect of the CGU algorithm, and is also essential in determining the global underestimating function $\Psi(\phi)$ as given by (3) and (4). The distribution of local minima, in effect, represents the energy surface $F(\phi)$, and therefore is the most convenient way to visualize the energy landscape.

To further simplify the visualization of the energy landscape, we now show how it can be represented in a novel two-dimensional plot. To do this we represent the deviation of ϕ from ϕ_G in a suitable manner. We define the Root Mean Square Weighted Deviation (RMSWD) by

$$\overline{\Delta\phi} = \sqrt{[(\phi - \phi_G)^T D(\phi - \phi_G)]}. \tag{5}$$

We then use $\overline{\Delta\phi}$ as the horizontal coordinate of a plot, with the energy difference from F_G as the vertical coordinate. Specifically, we have

$$\Psi(\phi) - F_G = \frac{1}{2}(\overline{\Delta\phi})^2 \tag{6}$$

so that plotting $\Psi(\phi)$ and F_j, j=1,...,k, vs. $\Delta\phi$ shows the energy landscape and its relationship to the underestimating energy surface. The global minimum conformation ϕ_G, and its corresponding energy $\Psi(\phi_G) = F_G$ are shown on the plot as a point at the origin. Those local minima which lie on the surface, and those above it, can also be clearly distinguished. This is illustrated in Figure 3. The energy gap between the global minimum F_G and all other local minima F_j, j=1,...,k, is shown by the vertical distances of the points representing the local minima.

Based on the computational results for small proteins (described in the next section), we are now able to compute the global minimum conformation ϕ_G corresponding to a function $F(\phi)$ for each specific protein with up to 50 residues. However, the conformation ϕ_G computed in this way will not be the same as ϕ_N, the known native conformation for the same protein. Improved energy functions $F(\phi)$ are needed in order to accomplish this.

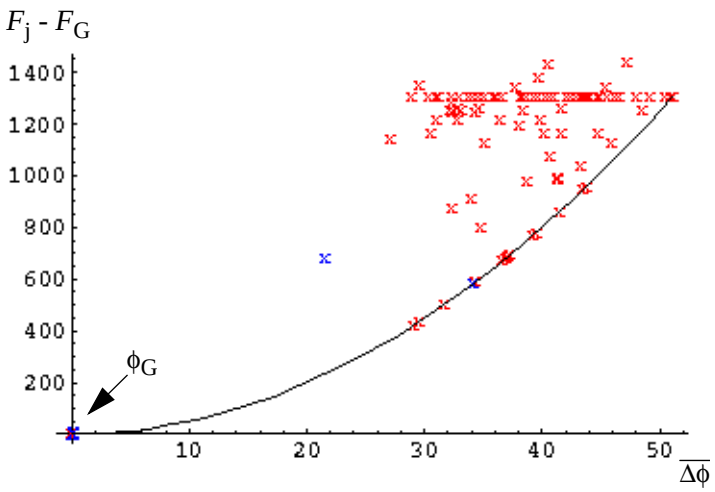

Figure 3 Example RMSWD Energy Landscape Projection Obtained from a 30-Residue Peptide Sequence

We can, however, use the knowledge of ϕ_N to modify $F(\phi)$ in a simple but non-physical way so that the modified $F(\phi)$ (denoted $\overline{F}(\phi)$) has its global minimum when $\phi = \phi_N$. This requires only a simple shift of coordinates to give the modified function

$$\overline{F}(\phi) = F(\phi - (\phi_N - \phi_G)) . \tag{7}$$

Clearly $\overline{F}(\phi)$ attains its global minimum value $\overline{F}_G$ at $\phi = \phi_N$. Therefore the shifted energy function $\overline{F}(\phi)$ attains the known native state ϕ_N at its global minimum $\overline{F}_G$ so this provides a shifted energy landscape for the corresponding protein molecule. It should be noted however that we must know ϕ_N to construct $\overline{F}(\phi)$, so that this does not solve the native structure prediction problem.

4. Physical Interpretation of the Global Underestimator Coefficients

The CGU method gives a very simple expression for the fluctuations around the native structure. The probability $P(\phi^{(j)})$ of finding a molecule in conformation $\phi^{(j)}$ is given by the Boltzmann distribution law

$$P(\phi^{(j)}) = \frac{e^{-(F_j - F_G)/k_B T}}{\sum_{i=0}^{N} e^{-(F_i - F_G)/k_B T}} \tag{8}$$

where k_BT is Boltzmann's constant multiplied by temperature, and k is the total number of conformations that are local minima of $F(\phi)$. Thus, higher energy states are less probable than lower energy ones.

Note that the CGU method will find only k of the N total local minima of $F(\phi)$, and that k « N is expected. However, for those N-k local minima not found, the corresponding energies are also expected to satisfy $F_j \gg F_G$, so that their effect on the total sum in (8) is negligible.

If $(\phi^{(j)})_i$ represents the i^{th} angle in the j^{th} conformation, then the weighted mean of the i^{th} angle is given by

$$\langle \phi_i \rangle = \sum_{j=0}^{N} P(\phi^{(j)}) \cdot (\phi^{(j)})_i$$

and the corresponding mean square deviation in $<\phi_i>$ is given by

$$\langle [\phi_i - \langle \phi_i \rangle]^2 \rangle = \sum_{j=0}^{N} P(\phi^{(j)}) \cdot [(\phi^{(j)})_i - \langle \phi_i \rangle]^2 \quad .$$

Thus a small mean square deviation demonstrates the increased reliability of $<\phi_i>$. Also a small mean square deviation should give $<\phi_i> \approx (\phi_G)_i$. If all such mean square deviations are small, then the computed global minimum angles $(\phi_G)_i$ should give a good approximation to the true native conformation $(\phi_N)_i$.

Since the final convex global underestimator $\Psi(\phi)$ agrees with the global minimum potential energy F_G at the computed global minimum conformation ϕ_G, then as stated in (3)

$$\Psi(\phi) - F_G = \frac{1}{2} \sum_{i=1}^{n} d_i [\phi_i - (\phi_G)_i]^2 .$$

Now, if ϕ' denotes a conformation with all angles ϕ_i, except for ϕ_l, fixed at their respective global minimum values $(\phi_G)_i$, then the energy difference directly attributed to any ϕ_l is clearly

$$\Psi(\phi') - F_G = \frac{1}{2} d_l [\phi_l - (\phi_G)_l]^2 . \tag{9}$$

Finally, with a suitable assumption[1], by applying (9) to (8), the Boltzmann distribution of angle ϕ_l is then proportional to (ignoring the denominator in (8))

---

1 For each conformation $\phi^{(j)}$, the CGU function value $\Psi(\phi^{(j)})$ matches the corresponding potential energy function $F_j \equiv F(\phi^{(j)})$. Even if this assumption is not satisfied, an upper bound on the standard deviation given in (10) may be obtained.

$$P(\phi') = e^{-\frac{1}{2}d_l[\phi_l - (\phi_G)_l]^2/k_BT} = e^{\frac{-1}{2\sigma_l^2}[\phi_l - \bar{\phi}_l]^2}$$

where $\bar{\phi}_l$ is the mean, and σ_l^2 is the variance. Therefore, we can interpret $(\phi_G)_l$ as the mean value of ϕ_l, and k_BT/d_l as the variance of ϕ_l obtained directly from the convex global underestimator. Note that a large value of d_l implies a small variance in the angle ϕ_l. Also a high temperature T, as well as a small value of d_l, implies a large variance in the angle, as expected. The standard deviation of ϕ_l is

$$\sigma_l = (k_BT/d_l)^{1/2}. \tag{10}$$

Note again that this result depends on the property that the CGU algorithm computes a large set of local minima, in addition to the global minimum.

5. Computational Results

We have studied small peptide sequences ranging in size from 5 residues to 36 residues. These sample peptides are: 5 residue met-enkephalin (MET), 9 residue oxytocin (1XY1), a 23 residue beta-beta-alpha motif [8] (BBA1), a 30 residue zinc-finger (7ZNF), and 36 residue avian pancreatic polypeptide (1PPT). In some cases the native structures for these peptides are also known (1XY1, 7ZNF, and 1PPT). Table 1 shows the results obtained from applying the CGU method on a 32 processor Cray T3E at the San Diego Supercomputer Center.

Table 1 CGU Results for Five Small Peptide Sequences on a 32 Node Cray T3E

Compound	Residues	CGU Native Energy	Solution Time
MET	5	-178.11 kcal/mol	6 s
1XY1	9	-355.02 kcal/mol	64 s
BBA1	23	-1424.96 kcal/mol	14 m
7ZNF	30	-1302.67 kcal/mol	53 m
1PPT	36	-2332.64 kcal/mol	2.4 h

The method does not correctly predict the known native structures, but this is not our goal here. The search strategy always finds conformations lower in energy than the true native structure, indicating limits of the energy function, not the

search strategy. The lowest energy conformations in the model have properties common to true native structures: compact states with hydrophobic clusters and hydrogen bonded secondary structure. Hence we believe that the CGU method is efficiently finding global solutions using our simple model and energy function, but that the energy function requires improvement.

As an example, the known native structure for 7ZNF (zinc-finger motif) consists of a single "tight" alpha helix, a hair-pin turn, and then another hairpin turn (Figure 4), whereas the corresponding CGU computed structure shows only a

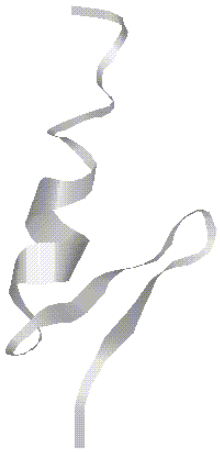

Figure 4 Known Native Structure for 7ZNF

slight tendency toward forming the first alpha helix, followed by a hairpin turn, and then two more hairpin turns (Figure 5). Figure 6 shows this same structure which highlights the effects of the very powerful attraction between the hydrophobic amino acid sidechains responsible for the compact clustering at the bottom of the figure. Clearly the hydrophobic attraction in this case dominates the hydrogen bond formation required for the construction of the tight alpha helix, and also contributes to the formation of the extra hairpin turn. We attribute this error not to the CGU search strategy, but instead to limitations in the accuracy of the potential function.

The CGU search strategy is finding global, or at least near global, solutions for the given model. Table 2 shows the probability distribution of all local minima obtained for each of the peptide sequences. For all test cases, the energy gap between the "best" computed structure and all other structures is so large that the probability that a peptide would be observed in the global minimum state is 1. In addition, the landscape projections, based on the RMSWD metric, such as the one shown in Figure 3, indicate that the landscape CGU is a tight underestimator for

Figure 5 CGU Computed Structure for 7ZNF

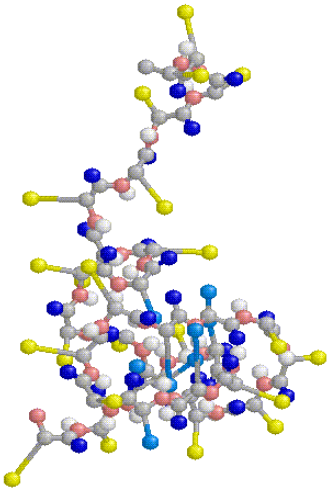

Figure 6 Ball and Stick Representation of the CGU Computed Structure for 7ZNF

many local minima and the energy gap between the global minimum and all others is significant.

Finally, based on the computational results obtained so far, the average running time, as a function of peptide length n, is $O(n^4)$ [3]. This function has been com-

Table 2 Probability Distribution for All Local Minima

	Number of Local Minima in Probability Range		
Compound	>.99	.99-.01	<.01
MET	1	0	5
1XY1	1	0	22
BBA1	1	0	62
7ZNF	1	0	93
1PPT	1	0	119

puted (based on the results presented in both this and the next section) to be $T(n) \approx (8.8e\text{-}5)n^4$ minutes for an n residue structure on a 32 node Cray T3E. Table 3 shows the exact values of T(n) for various values of n and Figure 7 shows a plot of

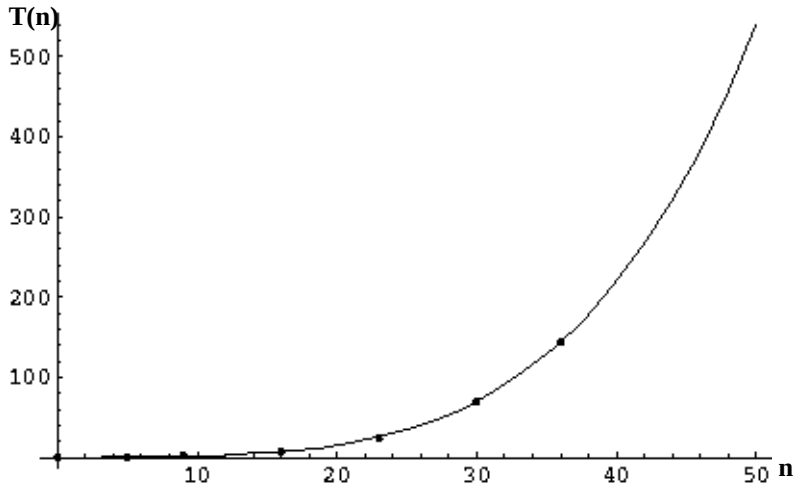

Figure 7 Solution Time, T(n), in Minutes as a Function of the Number of Residues n

T(n) along with the results from the five sample peptides.

Table 3 Average Running Time T(n) for Various n, on a 32 Node Cray T3E

n	10	20	30	40	50	100
T(n) minutes	2	15	70	220	539 (9 hrs)	8753 (6 days)

6. Effect of Sequence on Structure

The previous section shows that the average running time of the CGU algorithm is dependent on the chain length n of the peptide sequence. However, unlike many other search strategies, the running time of the CGU method is *independent* of the ordering of the residues within the sequence, and is even *independent* of the monomer composition. This is not to say that the global solutions obtained by the CGU method are unaffected by sequence variations. Indeed they are. However, the running time remains insensitive to the precise nature of the sequence.

Figures 8, 9, and 10 show the computed global minimum conformations for

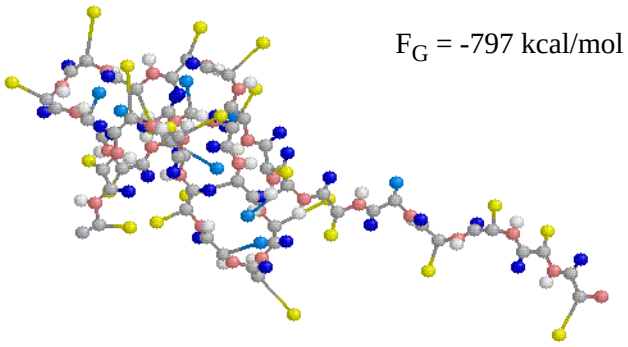

Figure 8 Global Minimum Structure of Permutation #1 of a 30 Residue Sequence

three permutations of the 7ZNF sequence (a 30 residue sequence). In each case, the same set of residues were permuted to provide a different ordering of those residues. Hence, the percent hydrophobicity remained constant in each case. From the figures, it is clear that the model native structures vary greatly, as would be expected with different monomer sequences. The solution times for these permuted test cases were 2430 s (requiring 3 iterations of the CGU method), 6293 s (requiring 4 iterations), and 2432 s (requiring 3 iterations). Hence the average time and the average time per iteration of the CGU method were 3718 s and 1065 s, respectively.

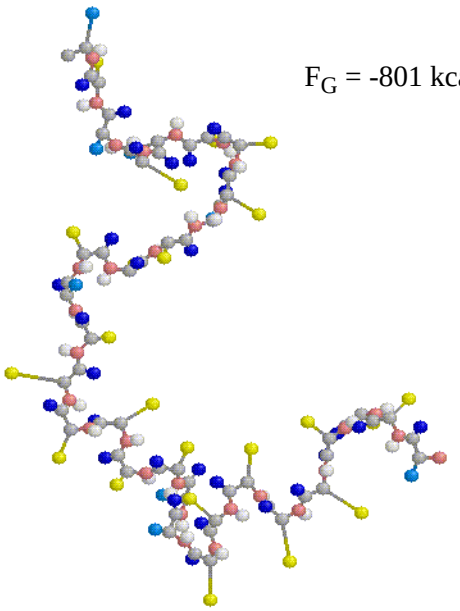

Figure 9 Global Minimum Structure of Permutation #2 of a 30 Residue Sequence

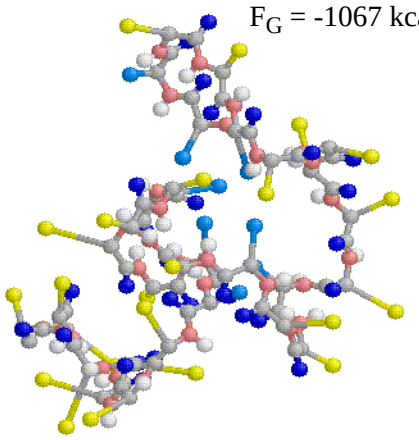

Figure 10 Global Minimum Structure of Permutation #3 of a 30 Residue Sequence

In total, five permutations each of sequences of lengths 5, 9, 16, 23, and 30 residues were tested. The results are shown in Tables 4, 5, 6, 7, and 8. In each case,

Table 4 5-Residue Permutation Results

Sequence	Time (s)	Iterations	Time/Iteration (s)	F_G
5mer #1	7	3	2.33	-178
5mer #2	15	7	2.14	-89
5mer #3	6	3	2.00	-89
5mer #4	5	3	1.67	-89
5mer #5	4	6	0.67	+0
Average	7.4	4.4	1.76	-409

Table 5 9-Residue Permutation Results

Sequence	Time (s)	Iterations	Time/Iteration (s)	F_G
9mer #1	125	3	41.67	-419
9mer #2	170	7	24.29	-267
9mer #3	50	4	12.50	-442
9mer #4	106	6	17.67	-470
9mer #5	207	3	69.00	-357
Average	132	4.6	33.03	-391

Table 6 16-Residue Permutation Results

Sequence	Time (s)	Iterations	Time/Iteration (s)	F_G
16mer #1	476	2	238	-89
16mer #2	402	3	134	-267
16mer #3	259	3	86	-178
16mer #4	834	7	119	-178
16mer #5	243	2	122	-178
Average	443	3.4	140	-178

Table 7 23-Residue Permutation Results

Sequence	Time (s)	Iterations	Time/Iteration (s)	F_G
23mer #1	848	2	424	-1588
23mer #2	1935	5	387	-1639
23mer #3	1290	4	323	-1560
23mer #4	1312	4	328	-1248
23mer #5	2027	5	405	-1640
Average	1428	4	373	-1535

Table 8 30-Residue Permutation Results

Sequence	Time (s)	Iterations	Time/Iteration (s)	F_G
30mer #1	2430	3	810	-797
30mer #2	6292	4	1573	-801
30mer #3	2432	3	811	-1067
30mer #4	3966	2	1983	-712
30mer #5	8424	9	936	-1062
Average	4223	4.2	1222	-888

there is clear variation in native structures for each permuted sequence. However, there is no clear dependence of computational time on the sequence.

In addition, the running time of the CGU method is invariant with respect to the residue types within a peptide sequence. In particular, for a given sequence of H and P type residues, the CGU algorithm is time invariant with respect to the specific H type residues and specific P type residues composing the sequence.

Table 9 shows the results of varying the H residues and P residues within a single

Table 9 23-Residue Permutations of H and P Residues Separately

Sequence	Time (s)	Iterations	Time/Iteration (s)	F_G
23HP #1	1328	3	443	-1723
23HP #2	2765	6	461	-1418
23HP #3	960	2	480	-1230
23HP #4	1037	3	346	-1156
23HP #5	1085	3	362	-1068
Average	1435	3.4	418	-1319

fixed sequence of length 23. In each case, the HP sequence was fixed, but the choices of H and P amino acids were randomly selected from the set of known H and P residues, respectively. Again, the time required to obtain the global minimum structure is invariant with respect to sequence. But again, the sequence itself affects the global solution greatly. In this case, since the ordering of the H and P residues remained fixed, but the types of H residues and P residues were varied, we would expect variation in the global solution due to the differences in the side chain volumes and the variable hydrophobic attraction expressed by each particu-

lar H residue. Figures 11 and 12 demonstrate this structural difference between

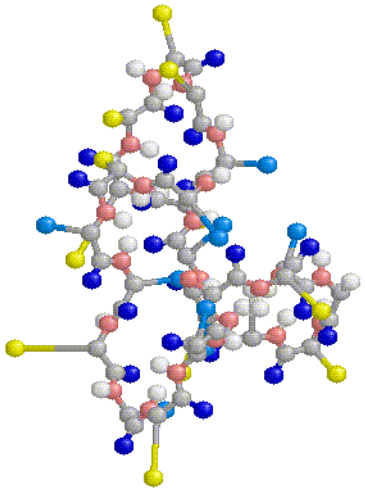

Figure 11 23-Residue Permutation #1 of H and P Independently

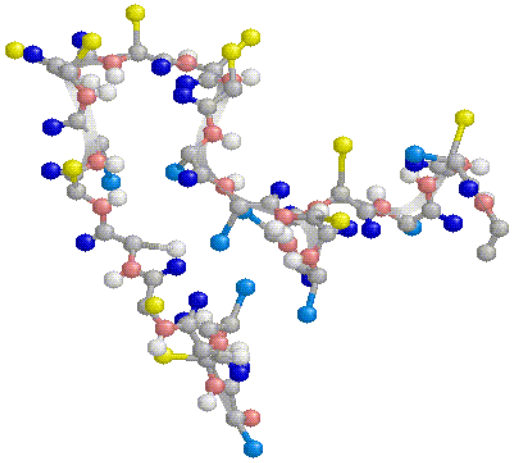

Figure 12 23-Residue Permutation #3 of H and P Independently

permutations 23HP #1 and 23HP #3.

7. Conclusions

The three principal results of this paper are: (1) to show that starting from many different randomly chosen open starting conformations of the chain, the CGU method converges on the same structure in each case, suggesting that the method is probably reaching the global minimum of the energy function, (2) to show the scaling of the solution time with the chain length, indicating that the method seems practical for small protein-sized molecules, and (3) to show that the computation time required by the CGU method is approximately independent of monomer sequence. The results also indicate that the energy landscapes are dominated by a single stable state which differs in energy by a wide energy gap from all other local energy states.

This paper is a test of a conformational search strategy, not an energy function. The energy function is not yet an accurate model of real proteins: the best computed structures differ from the true native structures. But similarly simple energy functions have begun to show value in predicting protein structures ([7], [10], and [11]). Therefore we believe improved energy functions used in conjunction with the CGU search method may be useful in protein folding algorithms.

Acknowledgments

K.A. Dill was supported by the NSF grant BIR-9119575, A.T. Phillips was supported by 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 **94**, Large Scale Optimization with Applications, Part III: Molecular Structure and Optimization (1997), L.T. Biegler et al. (Eds), 1-22.
3. K.A. Dill, A.T. Phillips, and J.B. Rosen, *Molecular Structure Prediction by Global Optimization*, Developments in Global Optimization (1997), I.M. Bomze et al. (Eds), 217-234.
4. K.A. Dill, A.T. Phillips, and J.B. Rosen, *Protein Structure Prediction and Potential Energy Landscape Analysis using Continuous Global Minimization*, Proceedings of the First Annual International Conference on Computational Molecular Biology, January 20-23, 1997, 109-117.
5. 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.
6. A.T. Phillips, J.B. Rosen, and V.H. Walke, *Molecular Structure Determination by Con-*

vex 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.

7. 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.
8. M.D. Struthers, R.P. Cheng, and B. Imperiali, *Design of a Monomeric 23-Residue Polypeptide with Defined Tertiary Structure*, Science **271** (1996), 342-345.
9. S. Sun, *Reduced representation model of protein structure prediction: statistical potential and genetic algorithms*, Protein Science **2** (1993), 762-785.
10. 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).
11. K. Yue, and K.A. Dill, *Folding Proteins with a Simple Energy Function and Extensive Conformational Searching*, Protein Science **5** (1996), 254-261.