# Simulation of Protein Folding

Rajiv Chopra, Ramesh Nagabushnam, Kripa Sundar and Paul Thadikaran
Department of Computer Science
State University of New York at Buffalo
Buffalo, NY 14260

## 1 The Problem.

The protein folding problem is one of the most important problems in biochemistry today. The problem is that of computing the three dimensional structure of a protein, given its amino acid sequence. Unger and Moult (1993) have demonstrated that computing the *native conformation* of a protein by modeling the folding process as a free energy minimization system is NP-hard (see also, Fraenkel, 1993).

We would like to work on an efficient parallel implementation of the modeling process in an effort to compute the *native conformation* of a given protein.

## 2 Motivation.

A protein is basically a chain of amino acids that are linked by peptide bonds. When a protein molecule gets created, it is essentially a (spatially) linear sequence of these amino acids. The molecule then folds into a complex three-dimensional *native conformation*, which determines its biological function.

Now, there exist efficient methods to determine the amino acid sequence of a given protein. However, the known methods of determining protein structure, namely, X-ray crystallography and Nuclear Magnetic Resonance (NMR) are very time consuming (Unger and Moult, 1993). Hence, an efficient method to compute the approximate structure from the known amino acid sequence of a protein is particularly relevant and important.

## 3 The Model.

The amino acid sequence contains all the information necessary for folding of the protein molecule. This is clear from the fact that proteins can be folded *in vitro*, without the aid of any other special folding mechanism or information. Each amino acid has a distinguished $C$-atom denoted by $C_\alpha$. The peptide bonds between amino acids are in fact bonds between the $C_\alpha$'s. Therefore, in modeling the folding of the protein molecule, it suffices to represent

it simply as a sequence of $C_\alpha$'s, ignoring the detail of the actual amino acids that constitute the protein.

We model the folding of the protein molecule by embedding its amino acid sequence in a three-dimensional cubic lattice. Each amino acid in the chain is represented by one element which can occupy one cubic cell of the lattice. The sequence is folded on the lattice such that two contiguous amino acids in the sequence must reside in two neighbouring cells of the lattice. This model was originally proposed by Unger and Moult.

It is clear that the number of possible foldings (*ie*, embeddings) of a chain on a lattice, which equals the number of non-self-intersecting walks on the lattice, is exponential on the length of the walk. Hence a brute-force approach will clearly not work on the search space of possible walks.

As Unger and Moult point out (p.1189) this model captures the essence of the protein folding problem. Holm and Sander (1991) have demonstrated that one can build a model of the full protein, once the coordinates of the $C_\alpha$ atoms are known. Further, it is known that the actual dihedral angles formed at the $C_\alpha$ "nodes" of the chain are constrained to a few pairs of values only (Ramakrishnan and Ramachandran, 1965; Herzberg and Moult, 1991). This is modeled by the restriction (by the model) of the possible directions of the chain at each $C_\alpha$ point to a few discrete values.

# 4 Problem Definition.

The protein folding problem can be viewed as a free energy minimization problem. The free energy function is defined on pairs of amino acids, and is dependent on their types and the distance between them. This gives us the following formalization of the protein folding problem:

**Input:**

- A sequence $S = \{s_1, s_2, \ldots, s_n\}$
- A cost function $c : S \times S \rightarrow Z^+$
- A distance metric $g : Z^3 \rightarrow R^+$

**Output:**

- A "folding" function, $f(s_i) \rightarrow Z^3$

**Constraints:**

- $\delta_x(f(s_i), f(s_{i+1})), \delta_y(f(s_i), f(s_{i+1})), \delta_z(f(s_i), f(s_{i+1})) \leq 1 \quad (0 < \text{i} < \text{n}) \quad$ *contiguity*
- $\delta_x(f(s_i), f(s_j)) + \delta_y(f(s_i), f(s_j)) + \delta_z(f(s_i), f(s_j)) \geq 1 \quad (1 \leq \text{i,j} \leq \text{n}) \quad$ *exclusion*
- $\text{E} = \Sigma_{i=1}^{n} \Sigma_{j \neq i} c(s_i, s_j) g(\delta_x(f(s_i), f(s_j)), \delta_y(f(s_i), f(s_j)), \delta_z(f(s_i), f(s_j)))$ is minimized,

where $\delta_x$, $\delta_y$, and $\delta_z$ refer to the respective axial distances of two points.

# 5 Possible Approaches.

That the problem is tough is clear from the fact that Unger and Moult express the opinion that even a simplification of the model to a two-dimensional lattice that only allows axial (*ie*, non-diagonal) chains is likely to be NP-hard. The reason for this difficulty seems to be the complexity of the energy hypersurface that characterizes the conformational states of biological macromolecules in general, and proteins in particular.

As Tuffery *et al* (1993) point out, the hypersurface typically has lots of energy minima that are often separated by barriers of widely differing heights. They have applied several optimization strategies to the problem of *protein side-chain conformation*, which exhibits most of the characteristics of the overall protein folding problem. Based on their studies, we conclude that a local optimization strategy based on the sparseness of the interaction matrix is likely to yield good results for our problem. We are also considering the simulated annealing approach, and would like to combine the two strategies in trying to improve the convergence figures for the *native conformation* computation. We expect that a machine like the CM-5 would provide a good platform for such an approach. Since we are still studying the possibilites, we would appreciate feedback from the DIMACS Committe on this issue.

We will be using a test-set of proteins whose crystallographic structures are already known. We plan to obtain this set ¿from the Protein Data Bank of Bernstein *et al* (1977).

# 6 References.


1. Bernstein, FC., TF. Koetzle, GJB. Williams, EF. Meyer, MD. Brice, JR. Ridgers, O. Kennard, T. Shimanouchi and M. Tasumi, 1977; "The protein data bank: a computer-based archival file for macromolecular structures", *J. Mol. Bio.* **112**, 535-542.

2. Fraenkel, Aviezri S., 1993; "Complexity of protein folding", *Bull. Math. Bio.* **55-6**, 1199-1210.

3. Herzberg, O. and J. Moult, 1991; "Analysis of the steric strain in the poly-peptide backbone of proteins", *Proteins* **11**, 223-229.

4. Holm, L. and C. Sander, 1991; "Database algorithm for generating protein backbone and side-chain coordinates from a $C_\alpha$ trace", *J. Mol. Bio.* **218**, 183-194.

5. Ramakrishnan, C. and G.N. Ramachandran, 1965; "Stereochemical criteria for poly-peptide and protein chain conformation", *Biophy. J.* **5**, 909-933.

6. Tuffery, P., C. Etchebest, S. Hazout and R. Lavery, 1993; "A critical comparison of search algorithms applied to the optimization of side-chain conformations", *J. Comput. Chem.* **14-7**, 790-798.

7. Unger, R. and J. Moult, 1993; "Finding the lowest free energy conformation of a protein is an NP-hard problem", *Bull. Math. Bio.* **55-6**, 1183-1199.