Protein Folding in the Hydrophobic-Polar Model on the 3D Triangular Lattice

Scott E. Decatur*
sed@theory.lcs.mit.edu

Serafim Batzoglou†
serafim@theory.lcs.mit.edu

MIT Laboratory for Computer Science

Abstract

We consider the problem of determining the three-dimensional folding of a protein given its one-dimensional amino acid sequence. The model we use is based on the Hydrophobic-Polar (HP) model [1] on cubic lattices in which the goal is to find the fold with the maximum number of contacts between non-covalently linked hydrophobic amino acids. Hart and Istrail [3] give a 3/8 approximation algorithm for folding proteins on the cubic lattice. Since the cubic lattice exhibits the "parity" problem (described below), we instead consider folding proteins on a different lattice: the 3D triangular lattice. We show a simple linear time algorithm which achieves a 16/30 asymptotic approximation. We briefly sketch a more sophisticated (but still linear time) algorithm which achieves a 62/105 factor of optimal.

Introduction

A long standing problem in molecular biology is the task of determining the native three-dimensional structure of a protein when only given the sequence of amino acid residues which compose the protein chain. Due to the complexity of the protein folding problem, scientists have studied a variety of simplifications of the general problem. Dill [1] introduced one such model, called the *HP Model*, which has been studied extensively. (Dill *et.al.* [2] review work in the HP and related models.) The foundation of this model is the belief that the first-order driving force of protein folding is due to a "hydrophobic collapse" in which those residues which prefer to be shielded from water (hydrophobic or non-polar residues) are driven to the core of the protein, while those which interact more favorably with water (hydrophilic or polar residues) remain on the outside of the protein. The native fold of a protein is hypothesized to be the fold with lowest free energy, or equivalently, the fold with the smallest surface area of hydrophobic residues exposed to water or polar residues.

In the HP model, a protein is represented as a string of beads and each residue is a single bead, either hydrophobic or polar. A fold of the protein within a given lattice is simply a self-avoiding walk along the lattice in which each residue is placed at a lattice point and residues which are adjacent in the primary sequence (*i.e.* covalently linked) must be placed at adjacent points in the lattice. A contact between two residues is considered to be a *topological* contact if they are not covalently linked and there is an edge connecting the lattice points of the two residues. The *free energy* of a fold is defined to be $(-1)\times$(# of topological contacts between hydrophobic residues). To simplify the presentation, we instead refer to the *score* of a fold which we define as the absolute value of its free energy. The goal is to find a fold of the protein which has the highest possible score. It is not known whether or not this problem is NP-hard. Hart and Istrail [3] give an algorithm for generating folds on a cubic lattice such that for any protein, the score of their fold is at least $3/8$ of the optimal score for this protein on a cubic lattice.

One major drawback of the cubic lattice is that if two residues are at any even distance from one another in the primary sequence then they cannot be in topological contact with one another when the protein is embedded in this lattice. We refer to this as the "parity" problem. Although the folds constructed by Hart and Istrail are at least $3/8$ of the optimal on the cubic lattice, the optimal on this lattice may by arbitrarily worse than the optimal on lattices without the parity problem. We therefore examine the HP folding problem on a regular lattice which does not exhibit a parity problem: the 3D triangular lattice.

The 3D Triangular Lattice

The lattice we use is shown in Figure 1 and is based on the β form of Silicon Carbide. The nodes in this lattice are defined to be the silicon atoms in the Silicon Carbide crystal. Two nodes in this lattice are connected if there exists a carbon atom that is bonded to both of the two silicon atoms corresponding to the nodes.

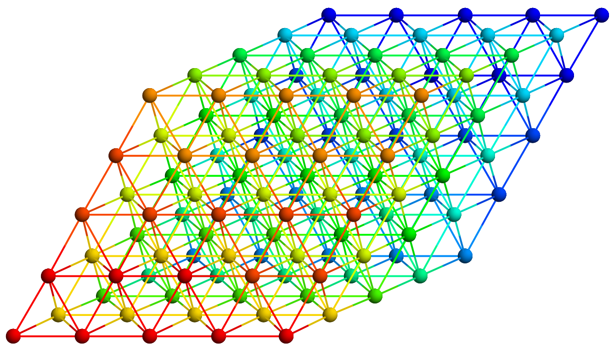

Figure 1: The Three-Dimensional Triangular Lattice.

It is not hard to verify that this lattice does not exhibit the parity problem. That is, for any two residues of distance at least two from one another, one can construct a fold of the protein along the lattice which has a non-covalent contact between these two residues.

*Scott Decatur is a postdoctoral researcher advised by Bonnie Berger and is supported by a grant from the Reed Foundation through the MIT School of Science and by ARPA Contract N00014-95-1-1246. Address: 545 Technology Square, Room 313, Cambridge MA 02139. Web: http://theory.lcs.mit.edu/~sed

†Serafim Batzoglou is a graduate student advised by Bonnie Berger and is supported by ARPA Contract N00014-95-1-1246. Address: 545 Technology Square, Room 342, Cambridge MA 02139.

Bounding the Optimal

In this lattice, every node has 12 neighbors. When embedding a protein in this lattice, 2 of these 12 neighbors must be the covalently linked residues.[1] Thus, each residue can be involved in at most 10 hydrophobic-hydrophobic contacts. Since each contact is shared by two residues, each residue may therefore contribute at most 5 hydrophobic-hydrophobic contacts. Thus, the optimal number of contacts on this lattice for a given protein is at most 5 times the number of hydrophobic residues.

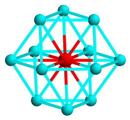

Figure 2: A single node and its twelve neighbors.

The Construction

The fold we construct places every hydrophobic residue in a single tightly-packed hydrophobic core. The core is composed of six residues at each level and is as deep as necessary to accommodate the entire protein. In Figure 3, a core of depth four is shown in which we wrap the protein in a clockwise direction from front to back. Thick lines represent the covalent connections, while thin lines represent the remaining adjacencies in the lattice. The central dark residues are hydrophobic while the outer light residues are polar. The fold shown is for a protein which has exactly two polar residues between each hydrophobic residue. If instead there were more or less than two polar residues between a given pair of hydrophobic ones, then these intervening polar residues would be laid out along a ladder as shown in Figure 4. Note that there exist nonoverlapping ladders of this form which extend indefinitely from each ladder of size 2 shown in Figure 3. Thus, we may accommodate any number of intervening polar residues.

Asymptotic Analysis

The asymptotic analysis assumes that the protein contains arbitrarily many hydrophobic (H) residues in order to ignore the end effects of the hydrophobic core. In the constructed core, each group of six H residues has 9 contacts within the group and 13 contacts to the next group. At most 6 of these connections can be covalent bonds. Thus, each group of six H residues has at least $9 + 13 - 6 = 16$ contacts. As noted above, each H residue can give rise to at most 5 new contacts. Thus, each group of six H residues can contribute at most $6 \times 5 = 30$ contacts and the asymptotic approximation is therefore $16/30$.

[1] If the residue is either the first or last of the protein, then only 1 neighbor is covalently linked, but this is unimportant for the asymptotic analysis.

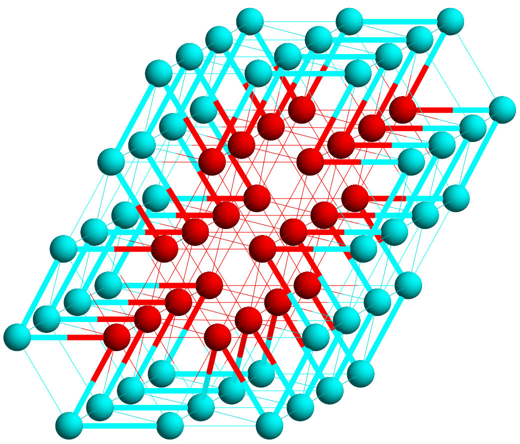

Figure 3: The 16/30 factor layout.

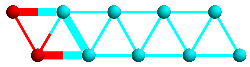

Figure 4: The ladder between hydrophobic i and $i + 1$.

Better Approximations

By altering the layout of the hydrophobic residues (not just the polar ones) as a function of the local "environment", one may achieve still better approximation ratios. We next sketch a construction which achieves a $62/105 \approx 3/5$ approximation.

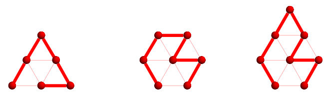

Figure 5: Fold types used in $62/105$ approximation.

The hydrophobic core in each level is one of three types, depending on the number of contacts that can be achieved among the residues of the same level. We construct the core incrementally, and we look at the next 6 H residues. If at least 3 of them have a P right neighbor, then they fold in the type I core. If more than 1 but less than 3 of the 6 residues have a P right neighbor, then these 6 H residues, together with the next (7th) H residue, fold in the type II core. If none of the 6 residues has a P right neighbor, then these together with the next two H residues fold in the type III core. There is an additional constraint, that a type III core cannot be on top of a type I core, and instead of that a type II core is used. Notice that in the case where type II or III cores are used, there always exist three consecutive H residues, so that the middle one of them is placed in the middle of the core. It is tedious but straightforward to verify that this construction can be realized, and that the worst case behavior is yielded by the sequence I-II-III-I-II-III-..., which gives an approximation to the optimal solution of at least $62/105$.

Acknowledgments

Thanks to Martin Farach, Jonathan King, Jon Kleinberg and Lior Pachter for discussion on protein folding and lattices. Images generated with the help of RasMol Molecular Renderer, Version 2.5.1 (Roger Sayle, October 1994).

References

[1] K. Dill. *Biochemistry*, 24:1501, 1985.

[2] K.A. Dill, S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, and H.S. Chan. Principles of protein folding: A perspective from simple exact models. *Prot. Sci.*, 4:561–602, 1995.

[3] W Hart and S. Istrail. Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal. In *Proceedings of the* 27^{th} *Annual ACM Symposium on the Theory of Computing*, 1995.