

Local Rules for Protein Folding on a Triangular Lattice and Generalized Hydrophobicity in the HP Model*

Richa Agarwala[†]
Martin Farach^{||}

Serafim Batzoglou[†]
Sridhar Hannenhalli^{**}

Vlado Dančik[§]
S. Muthukrishnan^{††}

Scott E. Decatur[¶]
Steven Skiena^{‡‡}

Abstract

A long standing problem in molecular biology is to determine the three-dimensional structure of a protein, given its amino acid sequence. A variety of simplifying models have been proposed abstracting only the “essential physical properties” of real proteins. In these models, the three dimensional space is often represented by a *lattice*. Residues which are adjacent in the primary sequence (*i.e.* covalently linked) must be placed at adjacent points in the lattice. A *conformation* of a protein is simply a self-avoiding walk along the lattice. The protein folding problem *STRING-FOLD* is that of finding a conformation of the protein sequence on the lattice such that the overall *energy* is minimized, for some reasonable definition of energy. This formulation leaves open the choices of a lattice and an energy function. Once these choices are made, one may then address the algorithmic complexity of optimizing the energy function for the lat-

tice. For a variety of such simple models, this minimization problem is in fact NP-hard [6, 5, 8].

In this paper, we consider the *Hydrophobic-Polar (HP) Model* introduced by Dill [2]. The HP model abstracts the problem by grouping the 20 amino acids into two classes: hydrophobic (or non-polar) residues and hydrophilic (or polar) residues. For concreteness, we will take our input to be a string from $\{H, P\}^+$, where P represents polar residues, and H represents hydrophobic residues. Dill *et al.* [1] survey the literature analyzing this model.

Selecting an energy function. Given a conformation, a pair of residues form a *topological contact* (or simply contact) if the residues are not covalently linked. A *bond* refers to a topological contact between a pair of H's. Define the *free energy* of a conformation as $(-1) \times (\# \text{ of bonds})$. The optimal conformation for the protein is the one which has the lowest free energy.

The biological foundation of this energy function 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 residues) are driven to the core of the protein, while those which interact more favorably with water (polar residues) remain on the outside of the protein. The protein is hypothesized to fold in such a way as to minimize the surface area of hydrophobic residues exposed to water or polar residues.

Selecting a lattice. HP simulations have typically followed Dill's original choice of square lattices. Unfortunately, one rather severe consequence of the structure of the square lattice is that no two amino acids can be at adjacent lattice points if the substring between them is of odd length. We call this the *parity constraint* of square lattices. Thus, the string $(PH)^n$ has no bonds in a square lattice, despite the fact that such a protein string has many potential bonds in “real space”.

The bizarreness of the parity constraint illustrates another possible pitfall of algorithmic analysis of this problem. An approximation ratio for a maximization algorithm is the ratio of a lower bound on the performance of the algorithm and an upper bound on the optimal solution. Thus, it is desirable to raise the lower bound of the algorithmic solution, or to lower the upper bound of the optimal solution. But in the case of the *STRING-FOLD* problem on a square lattice, the upper bound is artificially low, due to the parity constraint. To illustrate, consider once again the string $(PH)^n$, which has a trivial upper bound of 0 bonds, so any algorithm will achieve the optimum. Thus, any approximation ratio on a square lattice will have little meaning for moving towards a “realistic” solution of the problem.

*The extended abstract has appeared in the SODA'97 proceedings.

[†]National Center for Human Genome Research/National Institutes of Health, Bethesda, MD 20892, (*richa@helix.nih.gov*) under a contract with R.O.W. Sciences, Inc. Most of this work was done while this author was at DIMACS center, Rutgers University and was supported by Special Year National Science Foundation grant BIR-9412594.

[‡]MIT Laboratory for Computer Science and Department of Mathematics, 545 Technology Square, Room 342, Cambridge, MA 02139. (*serafim@theory.lcs.mit.edu*) Supported by a DOE Contract.

[§]Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113. (*dancik@hto.usc.edu*) Supported by NIH grant GM36230.

[¶]MIT Laboratory for Computer Science and Department of Mathematics, 545 Technology Square, Room 313, Cambridge, MA 02139. (*sed@theory.lcs.mit.edu*) Supported by a grant from the Reed Foundation through the MIT School of Science.

^{||}Department of Computer Science, Rutgers University, Piscataway, NJ 08855. (*farach@cs.rutgers.edu*) Supported by NSF Career Development Award CCR-9501942, an Alfred P. Sloan Research Fellowship, and NATO Grant CRG 960215.

^{**}Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113. (*hannenha@hto.usc.edu*) Supported by NSF Young Investigator Award, NIH grant 1R01 HG00987 and DOE grant DE-FG02-94ER61919.

^{††}Bell Laboratories, Lucent Technologies. Partly supported by DIMACS and the NATO grant CRG 960215.

^{‡‡}Department of Computer Science, State University of New York, Stony Brook, NY 11794-4400. (*skiena@cs.sunysb.edu*) Supported by ONR award 400x116yip01 and NSF Grant CCR-9625669.

Permission to make digital/hard copies of all or part of this material for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication and its date appear, and notice is given that copyright is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires specific permission and/or fee.

RECOMB 97, Santa Fe New Mexico USA

Copyright 1997 ACM 0-89791-882-8/97/01 ..\$3.50

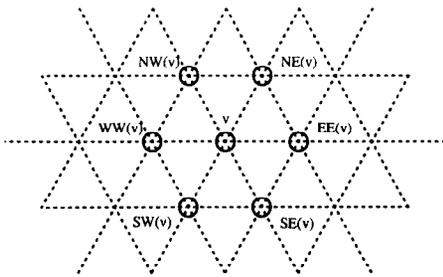


Figure 1: The two-dimensional triangular lattice

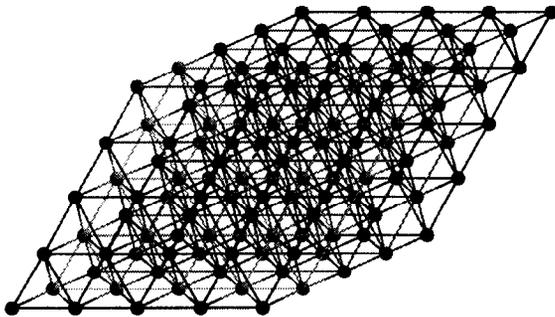


Figure 2: The Three-Dimensional Triangular Lattice.

If the square lattice is so seriously flawed, what is a more interesting lattice choice? We propose using triangular lattices as a folding model. The two- and three-dimensional triangular lattices are shown in Figure 1 and Figure 2. The triangular lattice does not exhibit the parity problem. That is, for every pair of sites x, y on any string, there exists a confirmation of the string on the triangular lattice such that x and y are neighboring sites on the lattice.

Past results. From a computational point of view, it is not known whether or not the protein folding problem under HP-model (on either square or triangular lattices) is NP-hard. The problem is known to be NP-complete when the alphabet size is unbounded and the lattice is a two- or three-dimensional square lattice [7]. Hart and Istrail [3] presented approximation algorithms on the square lattice having approximation factor of $1/4$ on the two-dimensional square lattice and $3/8$ in three-dimensions but the optimal conformations on these lattices may be arbitrarily worse than the optimal on lattices without the parity problem. It is therefore interesting to examine the HP folding problem on triangular lattices, and to strive for conformations approaching the more natural optimal score found there. The utility of the triangular lattice was also observed independently by Kleinberg [4].

Our results. We have developed a collection of local folding rules for the HP model on triangular lattices, in both two and three dimensions, and we prove approximation ratios for each of these rules. For all of our rules, these ratios are better than those achieved by Hart and Istrail [3] for the square lattice. As pointed out above, an approximation ratio can often be misleading. We provide these numbers as a rough guideline, since no more appropriate measure of goodness for a rule is available. Yet, since optimal solutions for triangular lattices are so much more densely packed than are optimal solutions for square lattices, we achieve our approximations by local rules which yield very dense

structures. That is, since the upper bound of the optimal solution is much higher, the performance of the folding algorithms must increase correspondingly. Furthermore, these local rules may be combined in many ways to get more complex foldings.

All of our folding rules are implementable in linear time. The following table summarizes the proposed rules for protein folding under the HP model in the two- and three-dimensional triangular lattices.

| Folding | 2D | 3D |
|------------------------|--------|---------|
| backbone | $1/4$ | $3/10$ |
| improved backbone | $1/2$ | $2/5$ |
| arrow | $1/2$ | $7/15$ |
| improved arrow | $6/11$ | n/a |
| star | n/a | $8/15$ |
| combined backbone-star | n/a | $44/75$ |
| improved star | n/a | $3/5$ |

We extend the HP model by considering a more general representation of hydrophobic residues. The new model is motivated by the fact that certain hydrophobic residues are more hydrophobic than others. While in the standard HP model all hydrophobic residues have identical hydrophobicities, our new model allows different residues to have different hydrophobicities and contacts between hydrophobic residues contribute to the energy function proportional to their combined hydrophobic strength. In the new model we are able to achieve similar constant factor approximation guarantees on the triangular lattice as were achieved in the standard HP model.

References

- [1] K. A. Dill, Sarina Bromberg, Kaizi Yue, Klaus M. Fiebig, David B. Yee, Paul D. Thomas, and hue Sun Chan. Principles of protein folding: A perspective from simple exact models. *Prot. Sci.*, 4:561–602, 1995.
- [2] K.A. Dill. Theory for the folding and stability of globular-proteins. *Biochemistry*, 24(6):1501–1509, 1985.
- [3] W. Hart and S. Istrail. Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal. *Journal of Computational Biology*, 3(1):53–96, 1996.
- [4] J. Kleinberg. Some computational problems in protein structure prediction. MIT EECS Area Exam (Cmte: R. Davis, T. Lozano-Perez, D.K. Gifford), December 1995.
- [5] R. Lathrop. The protein folding threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Engineering*, 7(9):1059–1068, 1994.
- [6] J. T. Ngo and J. Marks. Computational complexity of a problem in molecular structure prediction. *Protein Engineering*, 5(4):313–321, 1992.
- [7] M. Paterson and T. Przytycka. On the complexity of string folding. In *ICALP'96*, volume 1099 of *LNCS*, pages 658–669, 1996.
- [8] R. Unger and J. Moult. Finding the lowest free energy conformation of a protein is a NP-hard problem: Proof and implications. *Bulletin of Mathematical Biology*, 55(6):1183–1198, 1993.