

# On the Complexity of Protein Folding (Abstract)

Pierluigi Crescenzi

Deborah Goldman

Christos Papadimitriou

Antonio Piccolboni

Mihalis Yannakakis

Proteins are polymer chains consisting of monomers of twenty different kinds. Much of the genetic information in the DNA contains the sequence information of proteins, with three nucleotides encoding one monomer. In turn, proteins in an organism *fold* to form a very specific geometric pattern, known as the protein's *native state*. It is this geometric pattern that determines the macroscopic properties, behavior, and function of a protein. It is in general reasonably stable and unique.

The mapping from DNA sequences to monomer sequences is simple and very well-understood. In contrast, the mapping from the sequence of a protein to the geometric configuration of its native state—the “second half of the genetic code” [8]—is much more intricate and complex, and less understood; it has been the subject of intense investigation for decades. It seems clear that the forces underlying protein folding are the interactions between their monomers; recently, the view that *non-local* interactions dominate this process has been gaining ground [5]. To test this and other hypotheses concerning protein folding, researchers resorted to *simplified models* of proteins, mathematical abstractions of proteins that hide many aspects and exaggerate the effect of others; analysis and computer simulation of such models can then be compared to experimental results with actual proteins, to determine whether the emphasized aspects are indeed the dominant ones.

Perhaps the most successful and best-studied such model, and the one with apparently the best match with experiments<sup>1</sup>, is the *two-dimensional hydrophilic-hydrophobic model*, or H-P model, proposed by Dill [4]. In this model it is assumed that the protein is a sequence of 0s and 1s, and folding entails embedding the sequence in the two-dimensional lattice (see Figure 1). Each such folding is evaluated with a *score*, equal to the number of pairs of 1s that are adjacent in the lattice without being adjacent in the sequence; for example, in Figure 1 the score is five, corresponding to the five pairs of 1s connected by dotted lines. The score captures

<sup>1</sup>Chan and Dill [3] state that “for chain lengths for which exhaustive enumeration is possible (up to about 30 monomers), two-dimensional models more accurately represent the physically important surface-interior ratios of proteins than do three-dimensional models.”

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 98 New York NY USA

Copyright 1998 0-89791-976-9/98/3...\$5.00

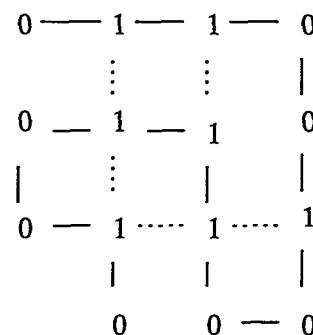


Figure 1: Embedding in the two-dimensional lattice.

a simple model of energy minimization, in which the “hydrophobic” 1s tend to be close to each other and thus avoid exposure, while 0s are neutral. It is assumed in this model that the native folded state is the one that maximizes score. It is therefore an interesting problem, given a sequence of 0s and 1s, to find the embedding on the lattice that maximizes score. Our main result is that *this problem is NP-complete*. See [2] for the full paper.

That proteins fold so as to minimize energy has been accepted for decades. This view quickly leads to a puzzling aspect of the problem, known as *Levinthal's paradox*, which can be paraphrased as follows “How can a folding protein choose so quickly among so many possible foldings the one with minimum energy?” [5]. Our result can thus be seen as a more compelling restatement of that paradox, since it implies that finding the optimum folding in the two-dimensional HP-model—the simplest abstraction of the protein folding problem one finds in the literature, and presumably a vast simplification of the true detailed 3-dimensional energy minimization problem in actual proteins—is NP-complete, that is to say, among the provably hardest problems of the sort alluded to by the paradox, in which we must optimize among an astronomical population of states.

There have been several NP-completeness results related to protein folding in the literature. A few years ago, several authors pointed out that certain general restatements of the problem, in which monomers attract or repel each other in ways that are general and can be used in encoding, are NP-complete [6, 10, 13]. More interestingly, it was proved in [11] that a combinatorial generalization of the H-P model to an infinite alphabet, of which one symbol is neutral like H-P's 0 symbol, and the score counts the number of adja-

cencies of elements with the same symbol, is NP-complete. More recently, [9] improved this to a finite, albeit very large alphabet. The present result is the first to settle the complexity of the simple two-dimensional H-P model actually proposed in the literature as the ultimate simplification of the protein folding problem. The H-P problem has been attacked from the point of view of approximation algorithms [7]; the present result sheds little light on this aspect of the problem, as our reduction is not in any interesting way approximation-preserving.

Our reduction is from the Hamilton cycle problem. As is common in previous proofs of weaker results, we start by showing that the folding problem for *sets of sequences* (that is, when many sequences are to be optimally folded) is NP-complete. We then proceed to establish the result for a single sequence, by resorting to certain interesting variants of the planar Hamilton cycle problem. In our proof we utilize an idea of Trevisan [12], whereby graphs can be embedded in the hypercube so that adjacency is captured by Hamming distance.

Our proof captures one of the basic intuitions of the H-P model, namely that hydrophobic monomers will tend to form a large "sphere" (in the two-dimensional lattice, a large hydrophobic square). Impurities in this sphere then must be aligned optimally to maximize score, and it is the complexity of this alignment that our proof captures. Finally, a version of our proof (in fact, without the planarity complication) can be used to prove the NP-completeness of the three-dimensional version of the protein folding problem in the H-P model —and in fact, the MAXSNP-completeness of the related problem of *minimizing losses* in three dimensions. Independently, and a few months prior to us, Berger and Leighton [1] proved that the three-dimensional protein folding problem in the H-P model is NP-complete; in fact, the approximability implications of their result for minimizing losses are stronger than ours.

## References

- [1] B. Berger, F. T. Leighton, manuscript submitted to *J. Mol. Biol.*, July 1997. Extended abstract appearing in this volume.
- [2] P. Crescenzi, D. G. Goldman, C. H. Papadimitriou, A. Piccolboni, M. Yannakakis, "On the complexity of protein folding," *Proceedings of STOC 1998*, to appear. For a full version, see <http://http.cs.berkeley.edu/~christos/hp.ps/>
- [3] H. S. Chan, K. A. Dill. The protein folding problem. *Physics Today* (1993), pp. 24-32.
- [4] K. A. Dill. Dominant forces in protein folding. *Biochemistry* 29 (1990), pp. 7133-7155.
- [5] K. A. Dill, S. Bromberg, K. Yue, K. Fiebig, D. P. Yee, P. D. Thomas, H. S. Chan. Principles of protein folding - A perspective from simple exact models. *Protein Science* 4 (1995), pp. 561-602.
- [6] A. S. Fraenkel. Complexity of protein folding. *Bulletin of Mathematical Biology* 55 (1993), pp. 1199-1210.
- [7] W. E. Hart, S. Istrail. Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal. *Proceedings of the 27th Annual ACM Symposium on the Theory of Computing*, 1995, pp. 157-168.
- [8] J. King. Deciphering the rules of protein folding. *Chemical Engineering News* 67 (1989), pp. 32-54.
- [9] A. Nayak, A. Sinclair, U. Zwick. Spatial Codes and the Hardness of String Folding. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, to appear.
- [10] J. T. Ngo, J. Marks, M. Karplus. Computational complexity, protein structure prediction, and the Levinthal paradox. In *The protein folding problem and tertiary structure prediction*, edited by K. M. Merz and S. M. Le Grand, Birkhauser, Boston, 1994.
- [11] M. Paterson, T. Przytycka. On the complexity of string folding. *Discrete Applied Mathematics* 71 (1996), pp. 217-230.
- [12] L. Trevisan. When Hamming Meets Euclid: The Approximability of Geometric TSP and MST [Extended Abstract]. *Proceedings of the 29th Annual ACM Symposium on the Theory of Computing*, 1997, pp.21-29.
- [13] R. Unger, J. Moult. Finding the lowest free energy conformation of a protein is an NP-hard problem: proof and implications. *Bulletin of Mathematical Biology* 55 (1993), pp. 1183-1198.