# Protein Folding in the Generalized Hydrophobic-Polar Model on the Triangular Lattice

Scott E. Decatur*
MIT Laboratory for Computer Science
sed@theory.lcs.mit.edu

May 1996

**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 [2] 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 [5] 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 three-dimensional triangular lattice. For this lattice, Decatur and Batzoglou [1] give a simple linear time algorithm which achieves a 16/30 asymptotic approximation. In this work, we further improve the model by generalizing the HP model to account for hydrophobic residues with different levels of hydrophobicity. After describing the motivation for this generalization of the model, we show that in new model we are able to achieve the same constant factor approximation guarantee on the triangular lattice as was achieved in the standard HP model.

KEYWORDS: Protein Folding, Lattice Models, Hydrophobicity Scales, Approximation Algorithms, Triangular Lattice

## 1 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 [2] introduced one such model, called the *Hydrophobic-Polar (HP) Model*.

*Supported by a grant from the Reed Foundation through the MIT School of Science. Address: 545 Technology Square, Room 313, Cambridge MA 02139. Web: http://theory.lcs.mit.edu/~sed

The HP model abstracts the problem by first grouping the 20 amino acids which compose proteins into two classes: hydrophobic (or non-polar) residues and hydrophilic (or polar) residues. In addition, the space in which the protein folds is discretized by defining a lattice and requiring residues to lie only on lattice points. Residues which are adjacent in the primary sequence (*i.e.* covalently linked) must be placed at adjacent points in the lattice. A fold of a protein is simply a self-avoiding walk along the lattice. In order to distinguish the quality of a given fold of a protein, we say a contact between two residues is a *topological* contact if they are not covalently linked and there is an edge connecting the lattice points of the two residues. Then the *free energy* of a fold is defined to be $(-1)\times$(# of topological contacts between pairs of hydrophobic residues). We often refer to the *score* of a fold which we define to be the absolute value of its free energy. The target fold for the protein is the one which has the lowest free energy, or equivalently, the highest possible score.

The biological 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 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. The HP model has been studied extensively and Dill *et.al.* [3] review much of the research in this and related models.

From a computational point of view, it is not known whether or not the problem of finding the fold with maximum score is NP-hard. Hart and Istrail [5] gave the first approximation algorithm for this problem. They gave an algorithm for generating folds on a cubic lattice (three-dimensional square lattice) such that for any protein, the score of the generated fold is at least $3/8$ of the optimal score for this protein on a cubic lattice.

Yet, a significant 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. It is therefore interesting to examine the HP folding problem on a regular lattice which does not exhibit the parity problem, such as the *triangular lattice*, and to strive for folds approaching the more natural optimal score found there. In Section 2, we describe the parity problem as well as the triangular lattices in more detail.

Decatur and Batzoglou [1] give a simple linear time algorithm for folding a protein on a three-dimensional triangular lattice which asymptotically guarantees a score at least $16/30 \approx 53\%$ of the optimal. This algorithm, which we later adapt, is presented in Section 3. The folds produced by this algorithm contain a helical hydrophobic core composed of repeated core planes each of which is formed by the exclusion of polar residues towards the outside of the core.

In this work, 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 in character than other hydrophobic residues. While in the standard HP model all hydrophobic amino acids have identical unit hydrophobic value, our new model allows different hydrophobic amino acids to have different integral hydrophobic values and contacts between hydrophobic residues contribute to the energy function proportional to their combined hydrophobic strength. We describe this motivation in more detail in Section 4 and then show that a simple variant of the algorithm of Section 3 can be used in the generalized HP model to achieve the same factor of approximation.

As mentioned above, optimal folding in the standard HP model is not known to be NP-hard and it might therefore be possible that the model is insufficient to capture the conjectured difficulty of real protein folding. As a step towards establishing the NP-hardness of this problem, a related model has been proposed by Paterson and Przytycka [7] in which there is an unlimited alphabet of residue types and only contacts between identical types make unit contribution to the energy function. In their model, they show that finding the optimal energy fold on a cubic lattice is NP-hard. Central to their proof is the use of arbitrarily many *different types* of residues and the rule that only exact matches between residues contribute. As these are not properties found in the folding of actual proteins, it would be preferable to show NP-hardness in a more plausible extension of the HP model. We hope that it will be possible to show that optimal folding in the generalized hydrophobic model of Section 4 is NP-hard and that therefore this more plausible model *does* in fact capture the conjectured difficulty of protein folding.

# 2 Lattices and the Parity Problem

As originally formulated, the HP model uses the square (2D) or cubic (3D) lattice. These lattices are simple to think about and to manipulate, but as mentioned above they posses a flaw referred to as the "parity problem" in which two residues of even distance from one another in the primary sequence are unable to be placed in contact with one another regardless of how one arranges the intervening sequence. For example, a protein sequence whose first residue is placed at the "*" in Panel A of Figure 1 may only have even numbered residues at "E" positions and odd numbered residues at "O" positions. An example sequence is shown in Panel B of Figure 1. Note that the cubic lattice also exhibits this problem.

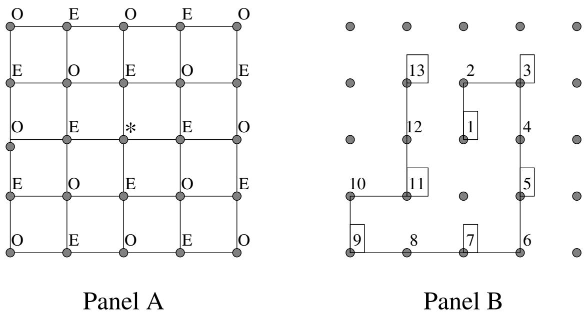


Figure 1: The Parity Problem.

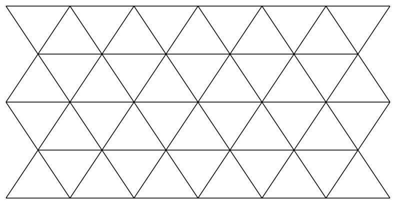

Figure 2: The Two-Dimensional Triangular Lattice.

This parity restriction is clearly *not* present when considering the real folding of proteins. For this reason, we instead consider protein folding in the HP model on a lattice which does not exhibit the parity problem, specifically the triangular lattice. The two-dimensional triangular lattice is shown in Figure 2. 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.

It is also important to note that the score of an optimal fold of a protein in the HP model can differ by an arbitrarily large amount between the square lattice and the triangular lattice. For example, when folding the sequence `HPHPHP...HP` (alternating hydrophobic and polar residues) on the square lattice, not even a *single* contact can be achieved between hydrophobic residues since all are at odd positions in the sequence. Yet, as shown in Figure 3, on the triangular lattice this sequence can be placed in a confirmation reminiscent of a protein beta-sheet which achieves a number of contacts between hydrophobic residues (dashed lines between solid circles) linear in the length of the sequence.

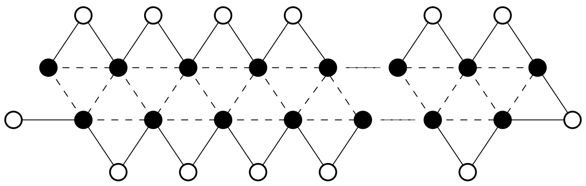

Figure 3: `HPHPHP...HP` folded on the triangular lattice.

As real protein folding occurs in *three-dimensional* space we in fact consider the lattice shown in Figure 4, which is a three-dimensional extension of the lattice shown in Figure 2.[1] It is not hard to verify that the three-dimensional lattice also does not exhibit the parity problem.

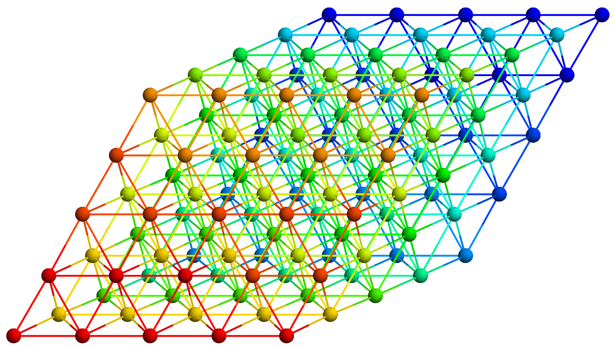

Figure 4: The Three-Dimensional Triangular Lattice.

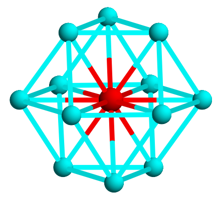

Figure 5: A single node and its twelve neighbors.

By examination of these triangular lattices, we may compute upper bounds on the score of a protein on such lattices in terms of the number of hydrophobic residues in the protein. In the two-dimensional lattice, each lattice point has six neighbors. Since each residue has two covalent neighbors, a residue at a lattice point may be in topological contact with at most four other residues.[2] Thus, each residue may be involved in at most 4 `H-H` contacts. Since each contact is shared by two residues, there may be at most 2 `H-H` contacts per hydrophobic residue, and therefore the optimal number of contacts on this lattice for a given protein is at most twice the number of hydrophobic residues in the protein. In the case of the three-dimensional triangular lattice, each node instead has 12 neighbors, as shown in Figure 5. Thus, the optimal number of contacts on this lattice for a given protein is at most $(12-2)/2 = 5$ times the number of hydrophobic residues in the protein.

[1]This three-dimensional lattice is based on the topology of the $\beta$ form of Silicon Carbide. The nodes in this lattice correspond to 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.

[2]If the residue is either the first or last of the protein, then only 1 neighbor is covalently linked. This may only increase the optimal score by 1 and is therefore ignored for our asymptotic analysis.

# 3 The Binary HP Construction on the 3D Triangular Lattice

The folding algorithm of Decatur and Batzoglou [1] for the three-dimensional triangular lattice places all hydrophobic residues in a single, tightly-packed, hydrophobic core. The core is composed of six residues at each level and contains as many levels as is necessary to accommodate all hydrophobic residues in the protein. In Figure 6, 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 hypothetical 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 8 in order to still place the next hydrophobic residue in its proper location. This allows the hydrophobic residues to be properly placed in the core regardless of the number of intervening polar residues. As shown in Figure 7, in each plane there exist nonoverlapping ladders of this form which extend indefinitely from each ladder of size 2 in Figure 6. Figure 9 shows how to switch from one plane to the next. When leaving position H6 in the front plane (dark), edge X is used to reach H1 in the next plane (light) if there are no intervening polar residues. Otherwise edge Y is used to reach position L1 at the entrance to the first ladder of the next plane.

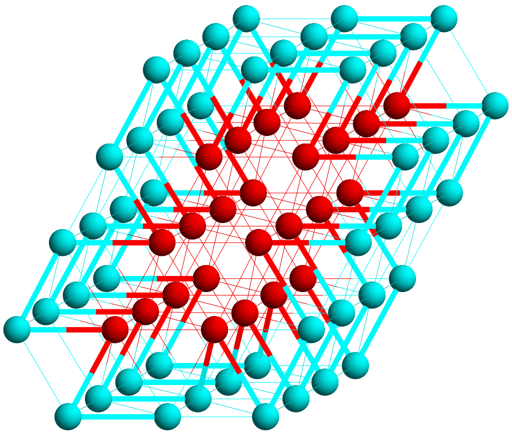

Figure 6: Four levels of the 16/30 factor layout.

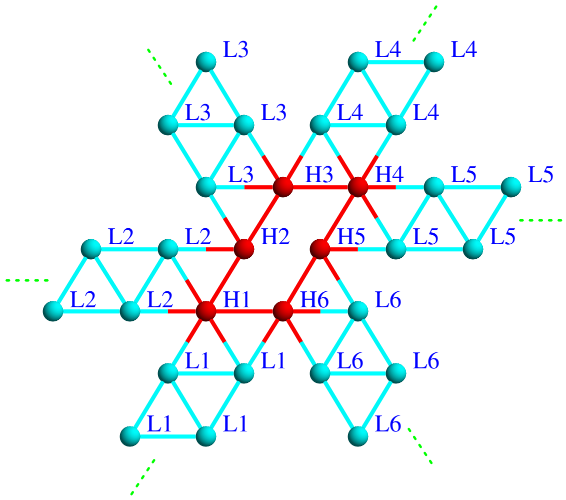


Figure 7: The six ladders in a plane.

The algorithm which folds protein sequences as described above can be clearly seen to run in linear time. In need only scan through the sequence once. Each hydrophobic is positioned in the next free location of the core. After a hydrophobic residue is positioned, the algorithm then counts the number of polar residues before the next hydrophobic and lays out the polar ladder of the correct size.

The asymptotic analysis assumes that the protein contains arbitrarily many hydrophobic residues in order to ignore the constant fewer contacts at the end of hydrophobic core. As shown in Figure 10, each group of six hydrophobic residues in the constructed core has 9 contacts within the group (thick lines) and 13 contacts to the next group (thin lines). At most 6 of these contacts could be covalent bonds. Thus, each group of six hydrophobic residues contributes at least $9 + 13 - 6 = 16$ topological

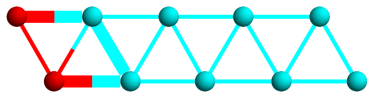

Figure 8: The ladder between hydrophobic $i$ and $i+1$.

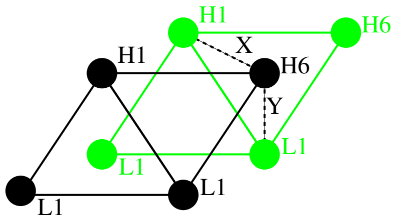


Figure 9: Switching levels.

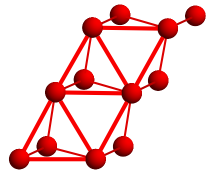

Figure 10: Hydrophobic contacts within one plane and to the next plane.

contacts. Therefore, the score of a protein folded in this manner asymptotically approaches at least $16\times$(# of hydrophobics)$/6$. As noted above, each hydrophobic residue can contribute at most 5 new contacts, and thus the asymptotic approximation is $(16/6)/5 = 16/30$.

# 4 Generalized Hydrophobicity

The standard HP model makes the simplifying assumption that all residues which are hydrophobic can have the same energy contribution to the hydrophobic collapse. Yet it is well known that different types of hydrophobic residues are more strongly hydrophobic than others. In fact, scientists have developed many such *hydrophobicity scales* in order to quantify the relative hydrophobicity of amino acids. (See for example Kyte and Doolittle [6] or Engelman, Steitz and Goldman [4].) We therefore propose an extension of the hydrophobic-polar model which accounts for residues of differing levels of hydrophobicity.[3]

In the new model, we allow each of the 20 amino acids to have a value from the set $\{0, 1, 2, \ldots\}$. Zero represents polar residues and non-zero values represent proportional levels of hydrophobicity. We then consider the value of a contact to be the amount of hydrophobicity buried from polar residues and solvent. Thus, a contact between any residue and solvent shields no hydrophobicity and is given a value of 0. A contact between a hydrophobic residue (with value from $\{1, 2, \ldots.\}$) and a polar residue (value 0) also shields no hydrophobicity and is given a value of 0. Finally, a contact between two hydrophobic residues (each with value from $\{1, 2, \ldots.\}$) shields both and is given value equal to their sum, the combined amount of hydrophobicity that is shielded.

If the ratio between the largest and smallest hydrophobic values assigned to the 20 amino acids is $\rho$, then blindly using an algorithm for the binary HP model could in the worst case result in a factor of $\rho$ loss in the approximation factor. In the case of the algorithm of Section 3, as $\rho$ grows arbitrarily the 16/30 approximation factor only falls by a constant to 2/5. These losses occur since some positions in a fold are involved in fewer contacts than others. If the strongly hydrophobic residues are placed in these positions, then the approximation suffers. Therefore, we must adapt our algorithms to adjust the construction based on the strength of the hydrophobic residues in the protein. Below, we adapt the algorithm of Section 3 for the generalized HP model and analyzed its performance. In addition,

[3]Note that although surface area of the actual residue may contribute to the strength of hydrophobicity attributed to a residue, we do not model the actual size differences in the spatial layout of the protein. Furthermore, we do not address the fact that polar residues also have diversity when compared to one another as well as when compared to water. These extension not considered here would still further improve the modeling of the protein folding problem.

other folding algorithms constructed for the binary HP model should also be examined in this respect and one would hope that their performance could also be kept comparable to their performance in the binary HP model.

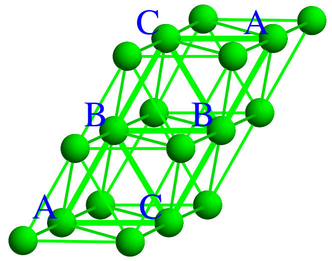


Figure 11: Contacts for positions A, B and C.

In adapting the algorithm described in Section 3, we are able to achieve the same 16/30 asymptotic approximation that was possible in the binary HP model. The new algorithm will once again place all hydrophobic (non-zero) residues in the core using the layout shown in Figure 6. In this layout, the number of contacts that a hydrophobic residue participates in depends on which of three different types of positions within the core it occupies (See Figure 11). If we number the hydrophobic residues starting at 0, then the number of hydrophobic contacts (2 of which may be covalent) in which the $r$-th hydrophobic residues participates is

$6 \Rightarrow$ if $r \equiv 0 \bmod 3$ (Position A)
$9 \Rightarrow$ if $r \equiv 1 \bmod 3$ (Position B)
$7 \Rightarrow$ if $r \equiv 2 \bmod 3$ (Position C)

out of a possible 12 contacts. Thus, there are three equivalence classes of residues, depending on their hydrophobic position in the primary sequence modulo 3. For $i \in \{0, 1, 2\}$, let $H_i$ be the set of residues at position $r$ such that $r \equiv i \bmod 3$, let $W_i$ be the sum of the hydrophobic values of all of the residues in set $H_i$ and let indices $i_1, i_2, i_3$ be such that $W_{i_1} \leq W_{i_2} \leq W_{i_3}$. If we define $W = W_{i_1} + W_{i_2} + W_{i_3}$, then the optimal score for a protein is at most $(12 - 2) * W$. Note that a single scan of the residues is sufficient to determine the values $W_0$, $W_1$ and $W_2$.

By placing the first hydrophobic residue at a different starting position in the construction, we effectively replace $r$ in the above equations with $r + 1$ or $r + 2$. Furthermore, by wrapping the core in the opposite direction, we effectively replace $r$ by $-r$. Therefore, we adapt the construction to place the residues of $H_{i_3}$ in position B, the residues of $H_{i_2}$ in position C, and the residues of $H_{i_1}$ in position A. If we let $x = W_{i_1}$, $y = W_{i_2} - W_{i_1}$ and $z = W_{i_3} - W_{i_2}$, then we achieve:

$$\begin{aligned}
\text{Score} &= (9-2) * W_{i_3} + (7-2) * W_{i_2} + (6-2) * W_{i_1} \\
&= 7 * W_{i_3} + 5 * W_{i_2} + 4 * W_{i_1} \\
&= 7 * (x + y + z) + 5 * (x + y) + 4 * (x) \\
&= 16x + 12y + 7z \\
&\geq 16x + \frac{32}{3}y + \frac{16}{3}z \qquad (1) \\
&= \frac{16}{3} \cdot (3x + 2y + z)
\end{aligned}$$

$$= \frac{16}{3} \cdot W$$

The approximation factor is therefore at least

$$\text{APPROX} \geq \frac{16/3 \cdot W}{10W} = \frac{16}{30}.$$

Inequality (1) is exact equality (yielding an approximation factor lower bound of $16/30$) when $y = z = 0$, *i.e.* $W_0 = W_1 = W_2$. When the weights of the three classes differ, the approximation factor bound increases and asymptotically approaches $7/10$ as all of the weight becomes concentrated in one of the three classes.

In some proteins, the weight of residues in one class (*e.g.* $W_0$) may be largest in the first half of the protein sequence, while the weight of residues in another class (*e.g.* $W_2$) may be largest in the second half of the protein sequence. In such cases, we would prefer one wrapping of the core in the first half and another wrapping in the second half. In order to change the wrapping in the middle of the sequence, there is a loss in the score due to the break in the core. Yet, in some cases this loss would be more than made up for by placing the heavier weighted residues at the B position in both halves of the sequence. Furthermore, this change in wrapping can be done as many times as is productive. Using dynamic programming, one could construct the best score possible by this strategy in time $O(nb)$ where $n$ is the number of residues in the protein and $b$ is the number of breaks permitted.

Note that these arguments for improving the approximation do not yield better *worst case* asymptotic approximations since here the worst case is when all classes have equal weight. But, since real proteins undoubtably have diverse hydrophobic makeup, the ability of an algorithm to leverage this diversity when present would be preferable. As this diversity is not guaranteed to exist in a worst-case analysis, it would also be of interest to determine if there are properties in biological protein sequences which HP folding algorithm *could* make use of to improve their performance guarantees, possibly using an average case analysis.

## Acknowledgments

Thanks to Bonnie Berger, 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] S. Decatur and S. Batzoglou. Protein folding in the Hydrophobic-Polar model on the 3D triangular lattice. In $6^{th}$ *Annual MIT Laboratory for Computer Science Student Workshop on Computing Technologies*, August 1996.

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

[3] 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.

[4] D.M. Engelman, T.A. Steitz, and A. Goldman. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Chem.*, 15:321–353, 1986. referenced from Branden and Tooze.

[5] 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.

[6] J Kyte and R.F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157:105–132, 1982. referenced from Branden and Tooze.

[7] M. Paterson and T. Przytycka. On the complexity of string folding. Submitted, April 1996.