INTERNATIONAL COMPUTER SCIENCE INSTITUTE
1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

Approximate Protein Folding in the HP Side Chain Model on Extended Cubic Lattices

Volker Heun

TR-98-029

December 1998

Abstract

One of the most important open problems in computational molecular biology is the prediction of the conformation of a protein based on its amino acid sequence. In this paper, we design approximation algorithms for structure prediction in the so-called HP side chain model. The major drawback of the standard HP side chain model is the bipartiteness of the cubic lattice. To eliminate this drawback, we introduce the extended cubic lattice which extends the cubic lattice by diagonals in the plane. For this lattice, we present two linear algorithms with approximation ratios of 59/70 and 37/42, respectively. The second algorithm is designed for a 'natural' subclass of proteins, which covers more than 99.5% of all sequenced proteins. This is the first time that a protein structure prediction algorithm is designed for a 'natural' subclass of all combinatorially possible sequences.

1 Introduction

One of the most important open problems in molecular biology is the prediction of the spatial conformation of a protein from its sequence of amino acids. The classical methods for structure analysis of proteins are X-ray crystallography and NMR-spectroscopy. Unfortunately, these techniques are too slow and complex for a structure analysis of a large number of proteins. On the other hand, due to the technological progress, the sequencing of proteins is relatively fast, simple, and inexpensive. Therefore, it becomes more and more important to develop efficient algorithms for determining the 3-dimensional structure of a protein based on its sequence of amino acids.

1.1 Amino Acids, Proteins, and Foldings

First we briefly review some basic facts on proteins from molecular biology. For a more detailed introduction, we refer the reader to fundamental textbooks on molecular biology and proteins (see, e.g, Hamaguchi [9]). A protein is a polymer built up from amino acids. An amino acid consists of a common main chain part and one of twenty residues which determines its characteristics. The chemical structure of an amino acid is illustrated in Figure 1. The main chain part consists of the central C-atom

$$\begin{array}{ccc} & H & \\ & | & \\ H_2N - & C_\alpha & - COOH \\ & | & \\ & R & \end{array}$$

Figure 1: Chemical Structure of an Amino Acid

(the so-called α-carbon C_α), an amino group (NH_2), a carboxy group ($COOH$), and a hydrogen atom. The residue R determines the chemical properties of an amino acid. In nature we find twenty different amino acids based on twenty different possible residues which can be as simple as a hydrogen atom in Alanine and as complex as two aromatic rings in Tryptophan.

A protein is a linear chain of amino acids linked together by peptide bonds. In a peptide bond the amino group and the carboxy group of two different amino acids are linked together by liberating a water molecule. The peptide bond between two amino acids is illustrated in Figure 2. Note that the six atoms in the shaded region in Figure 2 are located in a 2-dimensional plane. This follows from the interaction of the double bond of the carbonyl group (CO) and the non-binding pair of electrons in

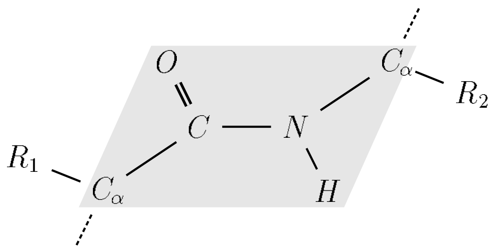

Figure 2: The Spatial Conformation of a Peptide-Bond

the nitrogen atom. This interaction implies that the bonds in C-O-N have a similar structure as the C-C bonds in benzene. Therefore, free rotation along the C-N bond is not possible due to the double bond character of the C-N bond. Furthermore, the configuration given in Figure 2, the so-called *trans* conformation, is more stable than the so-called *cis* configuration in which the positions of the hydrogen atom and the α-carbon adjacent to the nitrogen are swapped.

The sequence of amino acids for a given protein is called its primary structure. Driven by local interactions, short subsequences of the protein form individual spatial structures known as α-helices, β-strands, and reverse turns. Such conformations of subsequences are called the secondary structure of the protein. Based on global interactions of the amino acids, each natural protein folds into a unique spatial conformation called its tertiary structure. From the thermodynamic hypothesis it is assumed that the unique tertiary structure of a protein is the conformation with minimal free energy. Experiments have shown that the folding process *in vitro* is independent of external influence (by folding *in vivo* sometimes helper-molecules called chaperones are involved) and in general takes only a few seconds. It seems that the tertiary structure of a protein is encoded in its primary structure. Under this hypothesis, the spatial conformation of a protein may be computationally determined from its sequence of amino acids.

1.2 The HP Model

It is assumed that the hydrophobicity of amino acids is the main force for the development of a unique conformation. All natural proteins form one or more hydrophobic cores, i.e., the more hydrophobic amino acids are concentrated in compact cores whereas the more hydrophilic amino acids are located at the surface of the protein. This leads to a more simplified model, the so-called *HP model* (see, e.g., Dill [6] and Dill et al. [7]). Here, we distinguish only between two types of amino acids: hydrophobic (or non-polar) and hydrophilic (or polar). Therefore, a protein is modeled as a string over $\{H, P\}$, where each hydrophobic amino acid is represented by an H and each polar is represented by a P. In the following, a string in $\{H, P\}^*$ will also be called an *HP-sequence.*

The 3-dimensional space will be discretized by a cubic lattice. More formally, for $k \in \mathbb{N}$, let $\mathcal{L}_k$ be the following graph

$$\mathcal{L}_k = \left(\mathbb{Z}^3, \left\{\{x, x'\} \in \mathbb{Z}^3 \times \mathbb{Z}^3 \,\middle|\, |x - x'|_2 \leq \sqrt{k}\right\}\right),$$

where $|\cdot|_2$ is the usual Euclidean norm. Then $\mathcal{L}_1$ is the cubic lattice. A folding of a protein can be viewed as a self-avoiding path in the cubic lattice. More formally, a *folding* of an HP-sequence $\sigma = \sigma_1 \cdots \sigma_n$ is a one-to-one mapping

$$\varphi : [1:n] \to \mathcal{L}_k \quad \text{such that} \quad |\varphi(i-1) - \varphi(i)|_2 \leq \sqrt{k} \quad \text{for all} \quad i \in [2:n].$$

The *score* of a folding is the number of adjacent pairs of hydrophobic amino acids in the cubic lattice which are not adjacent in the given primary structure. Thus, the expected spatial conformation of a given protein is a folding with the largest score, since the negative score models the free energy. Therefore, a folding of a protein with the maximal score is called a *conformation.*

The major disadvantage of the the HP model is the representation of the 3-dimensional space by a cubic lattice because the cubic lattice is a bipartite graph. Thus, two hydrophobic amino acids with an even distance in the protein cannot contribute to the score, since they cannot be adjacent in the cubic lattice. In particular, all foldings of the sequence $(HP)^n$ are optimal, since each of its possible foldings on the cubic lattice has score 0. Therefore, we are interested in a more natural discretization of the 3-dimensional space. In this paper, we consider the *extended cubic lattice.* In the extended cubic lattice, we add to each lattice point 12 neighbors using diagonals in the plane, i.e., each lattice point has 18 neighbors. More formally, $\mathcal{L}_2$ is the mathematical description of the extended cubic lattice. Note that in $\mathcal{L}_2$ lattice points along a space diagonal are not connected.

A natural extension of the HP model is the *HP side chain model.* This is a more realistic model where the residues will be explicitly represented. In terms of graph theory, a protein is modeled as a caterpillar graph instead of a linear chain. More formally, a *caterpillar* of length n is the following graph $\mathcal{C} = (B \cup L, E)$, where

$$\begin{aligned} B &= \{b_1, \ldots, b_n\}, \\ L &= \{\ell_1, \ldots, \ell_n\}, \\ E &= \{(b_i, \ell_i) \mid i \in [1:n]\} \cup \{(b_{i-1}, b_i) \mid i \in [2:n]\}. \end{aligned}$$

Here, the set B represents the nodes in the *backbone* and L the so-called *legs.* A backbone node represents the C_α atom together with the main chain part of the amino acid whereas the leg represents its characteristic residue. This is still a simplification, since the residue can be as simple as a hydrogen atom in Alanine and as complex as two aromatic rings in Tryptophan. Note that we only mark the legs as hydrophobic or polar. Hence, a backbone node cannot increase the score of a folding.

1.3 Related Results

It is widely believed that the computational task of predicting the spatial structure of a given polymer (or, in particular, a protein) requires exponential time. First evidence for this assumption has been established by proving that the prediction of the conformation of a polymer for some more or less realistic combinatorial models is $\mathcal{NP}$-hard (see, e.g., Ngo and Marks [14], Unger and Moult [18], and Fraenkel [8]). For a comprehensive discussion of these lower bounds, we refer the reader to the survey of Ngo, Marks, and Karplus [15].

In [16], Paterson and Przytycka show that for an extended HP model with an infinite number of different hydrophobic amino acids it is $\mathcal{NP}$-hard to determine the conformation. In the extended HP model a protein will be modeled as a string over the (arbitrarily large) alphabet $(P, H_1, H_2, H_3, \ldots)$. Here only pairs of adjacent hydrophobic amino acids of the same type (i.e., contacts of the form H_i–H_i) contribute to the score. Recently, Nayak, Sinclair, and Zwick [13] improved this result. Even for a constant (but quite large) number of different types of amino acids the problem remains $\mathcal{NP}$-hard. Moreover, they also proved that this problem is hard to approximate by showing its MAXSNP-hardness. More recently, Crescenzi et al. [4] as well as Berger and Leighton [3] have shown independently that it is $\mathcal{NP}$-hard to determine the conformation the HP Model.

On the other hand, there is also progress on positive results on protein structure prediction. As a first milestone, Hart and Istrail exhibit in [10, 11] an approximation algorithm for protein folding reaching at least 3/8 of the optimal score in the HP model on the usual cubic lattice $\mathcal{L}_1$. In [12], the same authors present an approximation with a ratio of at least 2/5 in the HP side chain model on the cubic lattice.

In [2], Backofen presents an optimal algorithm for determining the conformation of a protein in the HP model on cubic lattices using constraint programming. Of course, this approach does not guarantee a polynomial running time of this algorithm. But nevertheless, it produces good results for short HP-sequences.

In [1], Agarwala et al. presented an algorithm with an approximation ratio of 3/5 for the HP model on the so-called triangular lattice (also known as face centered cubic lattice). This was the first approach to investigate non-bipartite lattices which has been initiated by Decatur [5]. Although the triangular lattice is differently defined, it can be topologically viewed as a superset of $\mathcal{L}_1$ and a subset of $\mathcal{L}_2$. An extension of the cubic lattice by just one plane diagonal direction in all three 2-dimensional subspaces is topologically isomorphic to the triangular lattice. Thus, in the triangular lattice each lattice point has 12 neighbors. Later, Hart and Istrail constructed in [12] a 31/36 approximation for the HP side chain model on triangular lattices. Note that the quality for all these approximation algorithms are measured with asymptotic approximation ratios.

1.4 Our Results

In this paper, we investigate protein folding on extended cubic lattices. The extended cubic lattice is a natural extension of the cubic lattice which bypasses its major drawback, its bipartiteness. First we present a general folding algorithm A which achieves for all protein sequences an approximation ratio of 59/70 ($\approx$84.3%). Then we describe a special folding algorithm B which can be applied to a restricted subset of HP-sequences. With the second algorithm we obtain an approximation ratio of 37/42 ($\approx$88, 1%). Although it is difficult to compare the approximation ratios for protein structure prediction algorithms on different lattice models, it should be mentioned that this is the best known approximation ratio for such algorithms.

Most of the known protein structure algorithms construct 'layered' foldings. This means that the algorithms constructs in reality a folding in the 2-dimensional sublattice from which the final folding in the 3-dimensional lattice will be generated. Therefore, only a few bonds use the third dimension. To obtain the high quality of the presented folding algorithm B, it is substantial to construct non-layered foldings in most parts of the conformation. Moreover, this construction does not only depend on the distribution of the hydrophobic amino acids in the protein as former algorithms. It also strongly depends on the length of contiguous subsequences of polar residues. This is strong evidence that the predicted folding is not too artificial.

On the other hand, this is the first time that folding algorithms for a 'natural' subclass of HP-sequences have been investigated. A strong indication that the considered subclass of HP-sequences is a 'natural choice' is the fact that more than 99.5% of all known sequences of proteins in the protein data base SWISS-PROT [19] (http://expasy.hcuge.ch/sprot/) belong to the considered subclass. Finally, the running time of both approximation algorithms are linear.

2 The General Folding Algorithm

In this section, we present a general folding algorithm in the HP side chain model on extended cubic lattices. Let $s{=}s_1 \cdots s_n$ be an HP-sequence. A sequence of HP-sequences $(\sigma_1, \ldots, \sigma_m)$ is called a *k-decomposition* of s iff the following four conditions hold:

1. $s = \sigma_1 \cdots \sigma_m$,
2. $\forall i \in [2, m{-}1] : |\sigma_i|_H = k$,
3. $0 < |\sigma_1|_H \leq k$ and $|\sigma_m|_H \leq k$,
4. the last symbol in each σ_i is an H for all $i{\in}[1{:}m{-}1]$.

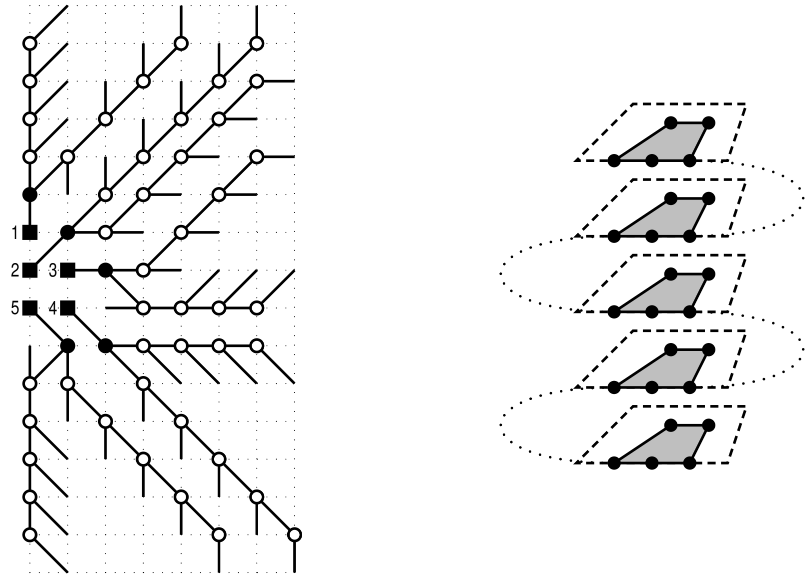

Figure 3: Folding of a Single 5-Fragment and Arrangement to a Pole

Here $|s|_H$ is the number of H's in the sequence s. The strings σ_i of a k-decomposition $(\sigma_1, \ldots, \sigma_m)$ are called *k-fragments*. If $|\sigma_1|_H = k$, we call σ the *canonical k-decomposition*. Note that for a given HP-sequence of length at least $2k-1$ there are k different k-decompositions.

Let s be an HP-sequence and let $\sigma=(\sigma_1, \ldots, \sigma_m)$ be the canonical 5-decomposition of s. First we fold each σ_i, for $i \in [1{:}m]$, as shown in the left part of Figure 3. Here, the nodes on the backbone of the protein are drawn as circles. More precisely, a backbone node is drawn black if it represents a hydrophobic amino acid and white otherwise. Hydrophobic residues are drawn as black squares, whereas the polar residues are not explicitly marked. The numbers in front of the squares represent the order of the hydrophobic residues in the sequence of amino acids. The contiguous block of polar amino acids between two hydrophobic amino acids are not connected in Figure 3. From the numbering of the hydrophobic residues, it should be clear which strands have to be connected and in which way.

We observe that for each 5-fragment consecutive backbone nodes with a hydrophobic residue are placed at neighbored lattice points, with the exception of the third and fourth backbone node. Therefore, the folding of the 5-fragment is still admissible even if the P-sequence between two hydrophobic residues is empty. If there is no polar residue between the third and fourth hydrophobic residue of a 5-fragment, we just remap the backbone node of the fourth hydrophobic residue one position up in the vertical direction.

In what follows, we show how to combine this folding of 5-fragments to obtain a folding in the 3-dimensional space. Using the third dimension, we combine the 5-fragments to a pole of height m such that the corresponding hydrophobic residues form a vertical column. This will be achieved by arranging the layers in a zig-zag-style in the third dimension which is sketched in the right part of Figure 3. Here only the hydrophobic residues are drawn explicitly as black circles. Note that at the front half of this pole the three hydrophobic residues have no neighbors outside the pole. Using a turn after $m/2$ layers, we combine the two halves to a new pole such that each layer contains 10 hydrophobic residues. A simple computation shows that each layer of 10 hydrophobic residues contributes 59 to the score: 23 hydrophobic-hydrophobic contacts within a layer and 36 H-H contacts to the two neighboring layers.

By definition, each lattice point has exactly 18 neighbors. Thus, each hydrophobic residue can have at most 17 contacts with other hydrophobic neighbors. This upper bound on the number of hydrophobic neighbors of a hydrophobic residue can be improved as follows. We denote by a *loss* an edge in the lattice with the property that a hydrophobic residue is mapped to exactly one of its endpoints. In the following, a backbone node with its adjacent leg is called a *basic pair*. A basic pair is called hydrophobic iff its leg is marked hydrophobic.

Lemma 1 *In all foldings on the extended cubic lattice, each single hydrophobic basic pair is incident to at least* 6 *losses.*

Proof: Consider a hydrophobic basic pair $(b,\ell)\in B\times L$. Assume that b and ℓ are mapped to adjacent lattice points p_b and p_ℓ, respectively. There exist at least 6 lattice points q_i, for $i\in[1{:}6]$ such that q_i is adjacent to both p_b and p_ℓ (for an illustration cf. Figure 4). Consider a fix (but arbitrary) lattice point q_i. Either a hydrophobic

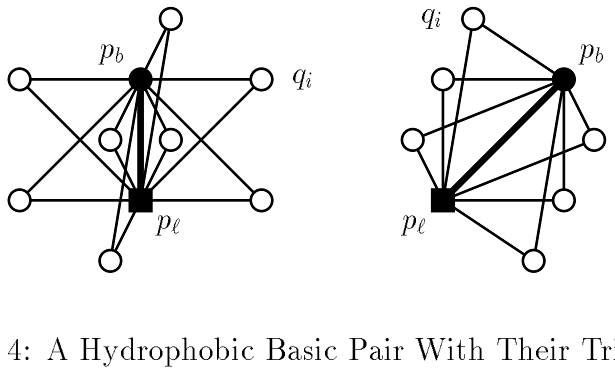

Figure 4: A Hydrophobic Basic Pair With Their Triangles

residue is assigned to q_i or not. In the first case, there is a loss along edge $\{p_b, q_i\}$; in the latter case, there is a loss along edge $\{p_\ell, q_i\}$. Hence, in both cases the hydrophobic basic pair is incident to a loss. ∎

Since at most two hydrophobic residues can be involved in each loss, we immediately obtain the following corollary.

Corollary 2 *In each folding on the extended cubic lattice, a hydrophobic residue is on average incident to at least* 3 *losses.*

Note that in general a single hydrophobic residue can have 17 hydrophobic neighbors. But in this case the neighbors have 3 additional losses, implying that on average each hydrophobic residue has at least 3 losses. Using Corollary 2, it follows that each hydrophobic residue can contribute to the score of a folding of at most $\frac{17-3}{2} = 7$. Our construction together with the previous lemma leads to the following theorem. Note that we consider asymptotic approximation ratios in this paper.

Theorem 3 *Algorithm A constructs a folding in the HP side chain model on extended cubic lattices for an arbitrary HP-sequence with an approximation ratio of at least* 59/70 *(*$\approx$84.3%*). Moreover, this folding can be computed in linear time.*

3 The Improved Folding Algorithm

In this section, we describe an improved folding algorithm B. This algorithm is designed for a special subset of HP-sequences. Intuitively, a protein sequence is perfect if each fragment has a one or two sufficiently short subsequences of contiguous polar residues.

Let s be an HP-sequence and let $\sigma{=}(\sigma_1, \ldots, \sigma_m)$ be a 6-decomposition of s. Further, let $\sigma_\nu{=}P^{\ell_1}H\cdots P^{\ell_6}H$ be 6-fragment. We call σ_ν *perfect* iff

(i) there exists $i{\in}[2{:}6]$ such that $\ell_i{=}0$, or

(ii) there exists $i{\neq}j{\in}[1{:}6]$ such that $\ell_i{+}\ell_j{\leq}3$.

An HP-sequence is called *perfect* if it has a 6-decomposition such that all its 6-fragments are perfect. If it has a 6-decomposition such that all but one of its 6-fragments are perfect, the HP-Sequence is called *nearly perfect*. The substrings P^{ℓ_i} for $i{\in}[1{:}6]$ are called an ℓ_i-block at position i. For example, the 6-fragment

$$\sigma{=}P^{27}HP^2HP^{11}HP^{12}HP^1HP^4H$$

is perfect and has a 12-block at position 4.

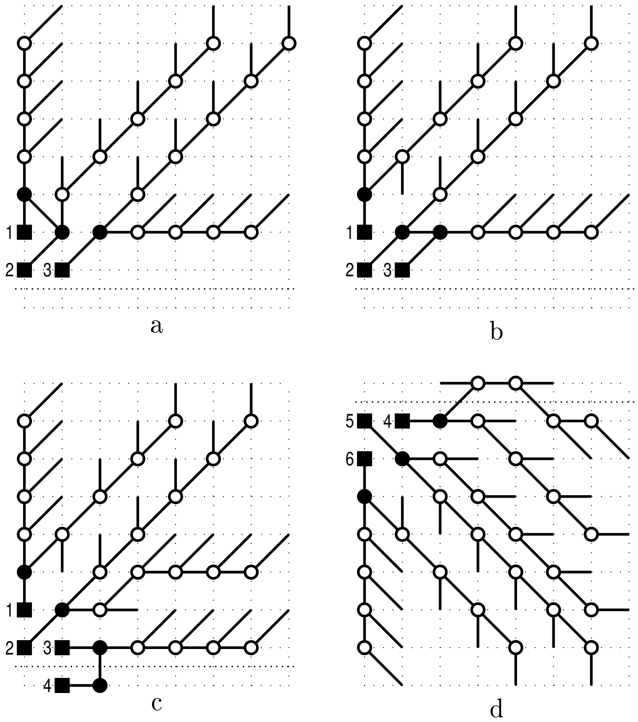

Figure 5: Folding of a 6-Fragment with a 0-Block (Case 1)

Again, we first describe how to fold a single 6-fragment. We will use two adjacent 2-dimensional planes to achieve the folding. In each plane, we will place 3 hydrophobic residues. We distinguish three cases depending on whether the 6-fragment is perfect because of a 0-block at position greater than 1, a combination of a 0-block at position 1 and a 3-block, or a combination of a 1- and a 2-block.

Case 1: First we assume that the 6-fragment is perfect because of a 0-block at position $i>1$. The folding is illustrated in Figure 5. In Figure 5a and 5b the foldings for a 6-fragment with a 0-block at position 2 and 3, respectively, are shown. The folding will be completed as illustrated in Figure 5d. In Figure 5c the first part of the folding of a 6-fragment with a 0-block at position 4 is shown. This folding will be completed by a reverse traversal of the same folding given in Figure 5c in the next layer. The case where the 0-block is at position 5 or 6 is symmetric to the cases where the 0 block is at position 2 or 1, respectively.

In contrast to the folding in the previous section, the folding of a 6-fragment consists of two layers with three hydrophobic residues each. In both layers the hydrophobic residues form a triangle. The narrow dotted horizontal lines in Figure 5 indicate where the 6-fragment will be folded to obtain this construction. □

Case 2: Now we consider the case of a 0-block at position 1. The Figures 6a, 6b, 6c, 6d, and 6e illustrate the folding in the cases where the 3-block is at position 2, 3, 4,

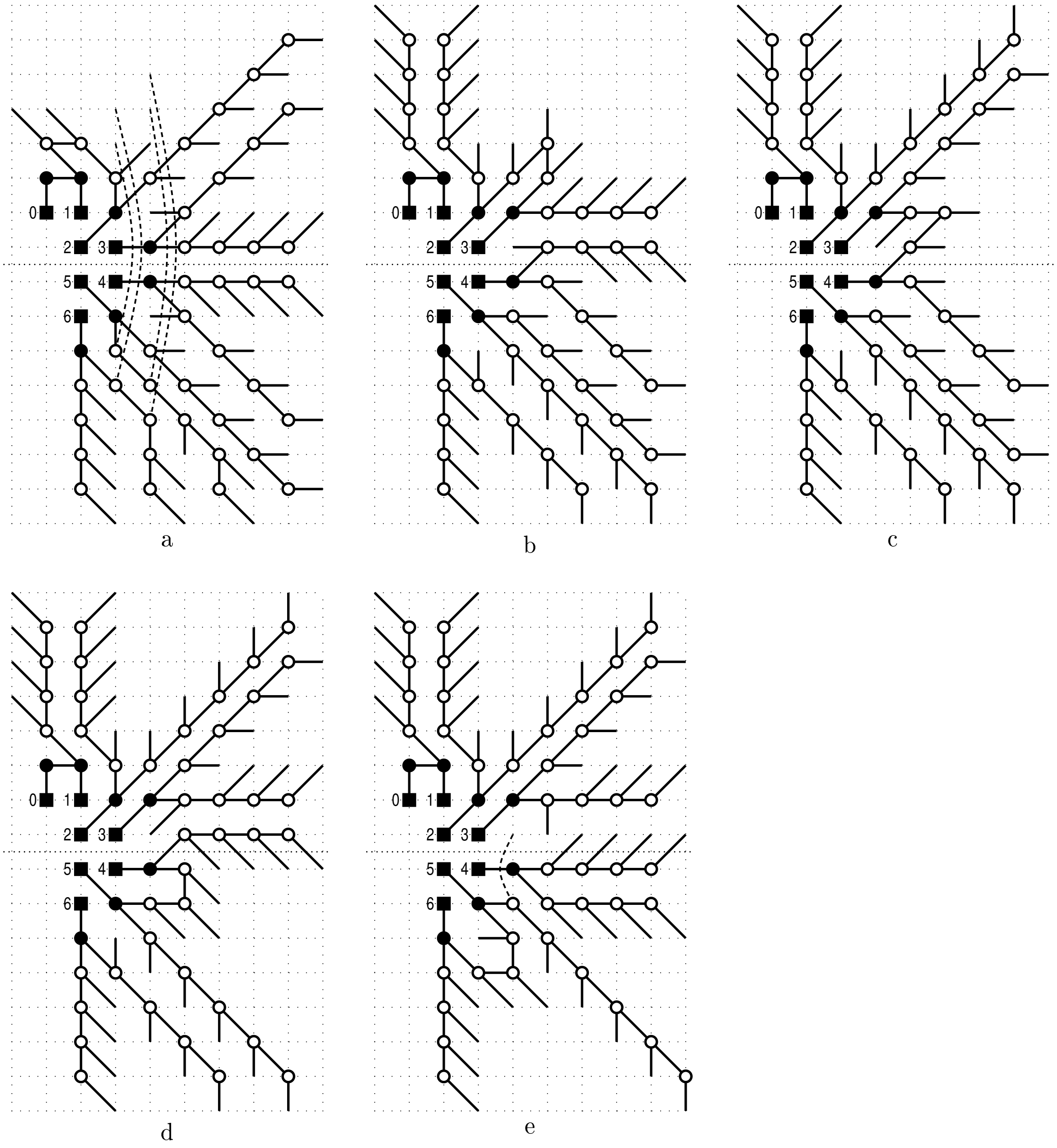

Figure 6: Folding of a 6-Fragment with a 0-Block at Position 1 and a 3-Block (Case 2)

5, and 6, respectively. Again, the narrow dotted horizontal line indicates where the folding will be folded to obtain two layers. The dashed lines indicates edges of the caterpillar which arise between adjacent layers.

Note that we use here some area which will be usually used to connect the last hydrophobic amino acid of the previously considered fragment with the first hydrophobic

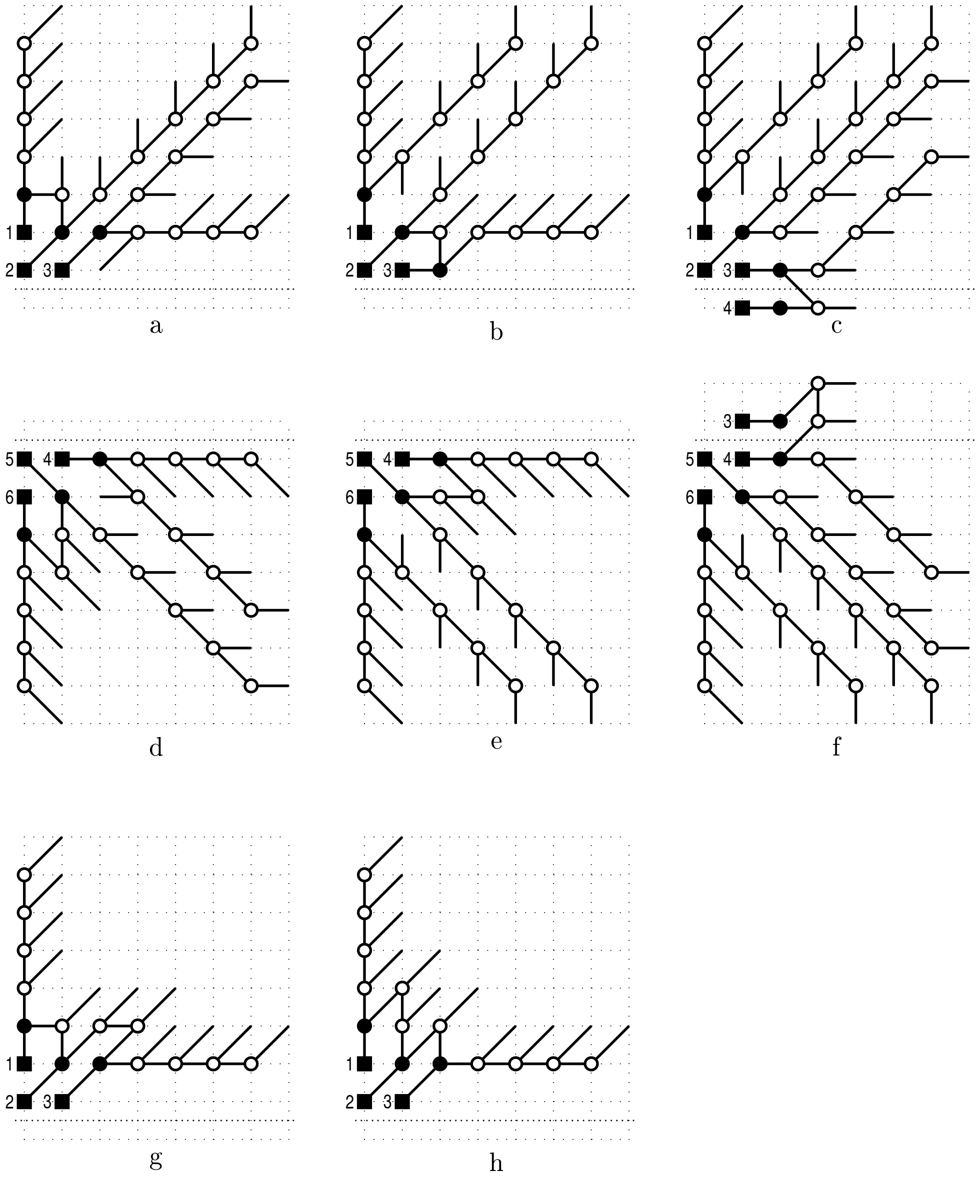

Figure 7: Folding of a 6-Fragment with a 1- and a 2-Block (Case 3.1)

residue of the actual fragment. In our construction, the used positions in the previous layer from the last visited hydrophobic residue are identical. Hence a reuse is possible and will not cause any difficulties. □

Case 3: Finally, we consider a combination of a 1- and a 2-block. Now we distinguish 3 subcases depending on whether at position 0 there is a k-block, a 1-block, or a 2-block for some $k>2$.

Case 3.1: The folding will be constructed from the partial foldings of a 6-fragment given in Figure 7 and Figure 5d. The following Table 1 shows how to combine these

1- \ 2-block	2	3	4	5	6
2	—	7g+5d	7a+7f	7a+7e	7a+7d
3	7h+5d	—	7b+7f	7b+7e	7b+7d
4	$(7a+7f)^R$	$(7b+7f)^R$	—	7c+7e	7c+7d
5	$(7a+7e)^R$	$(7b+7e)^R$	$(7c+7f)^R$	—	$(7h+5d)^R$
6	$(7a+7d)^R$	$(7b+7d)^R$	$(7c+7d)^R$	$(7g+5d)^R$	—

Table 1: Combinations of Subfoldings to a Folding of a 6-Fragment

partial foldings. The rows and columns refer to the positions of the 1- and 2-block, respectively. The superscript R indicates that the combined folding is traversed in reverse order.

For example, the folding of a 6-fragment of a 1-block at position 2 and a 2-block at position 5 is the combination of the foldings given in Figure 7a and Figure 7e.

Note that for the combination of Figure 7a with Figure 7f a minor modification of the folding given in Figure 7a is necessary. The backbone node of the hydrophobic residue labeled with 3 has to be remapped just to the right of the hydrophobic residue which is obviously possible. □

Case 3.2: Now we consider the case that the 1-block is at position 1 in the 6-fragment. Figure 8a illustrates the folding if the 2-block is at position 2. The folding for a 2-block at position 3 is obtained by a combination of the foldings given in Figure 8b and Figure 5d.

If the 2-block is at position 5 or 6, the folding will be combined from the foldings given in Figure 8c and Figure 7e or Figure 7d, respectively. If the 2-block is at position 4, the folding is more complex and illustrated in Figure 8d. Here, the dotted curves indicate connected subsequences of polar residues. Observe that the order of the traversed six hydrophobic is different from that in the other foldings. Here, the last visited node is directly above the fourth visited node of this fragment instead of the first one in the other foldings. □

Case 3.3: It remains the case where the 2-block is at position 1 in the 6-fragment. These are the most complex foldings and they are explicitly illustrated in Figures 9a through 9e depending on the position of 1-block. □

Note that all foldings are drawn for the case that the subsequences of contiguous polar residues may be arbitrarily long. But nevertheless our construction is also valid for any length of subsequences of contiguous polar residues with some minor modifications.

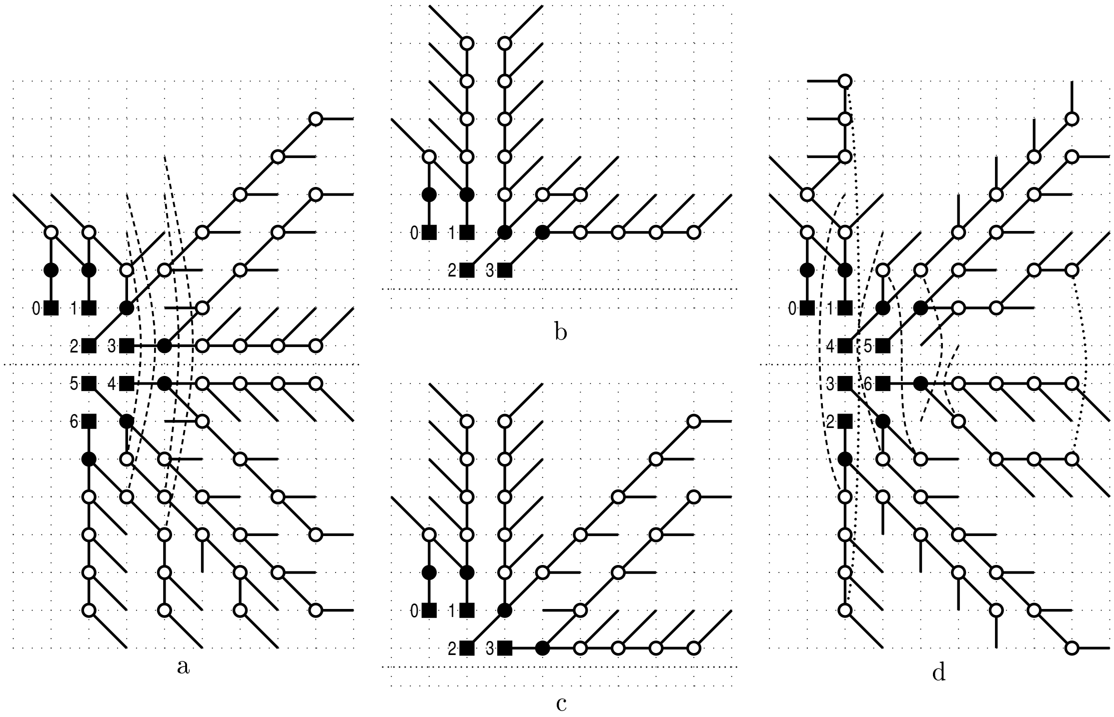

Figure 8: Folding of a 6-Fragment with a 1-Block at Position 1 and a 2-Block (Case 3.2)

It remains to construct a complete folding based on the presented foldings of the 6-fragments. First we combine the foldings of the 6-fragments to a long pole and break it into 4 parts $P_1, \ldots, P_4$ of equal height. Then the four parts will be arranged as shown in Figure 10. In Figure 10 only the hydrophobic residues are represented by gray quarters of the cylinder. For example, a folding of a single 6-fragment is illustrated in this figure by six black circles. The connections between these four quarters are drawn as dashed curves.

Note that the four parts can be arranged such that in P_1 and P_2 as well as in P_3 and P_4 the last and first hydrophobic residue are neighbors in the lattice. To connect the parts P_2 and P_3 we spent an extra 6-fragment because otherwise the contiguous block of polar residues between these parts may be too short. In the final folding, each layer consists of 12 hydrophobic residues. Each layer of 12 hydrophobic residues contributes 74 to the general score: 30 hydrophobic-hydrophobic contacts within a layer and 44 H-H contacts to the neighboring two layers. By Corollary 2, each layer can contribute on average at most 12∗7=84 to the score. Thus, we have proved the following theorem.

Theorem 4 *Algorithm B constructs a folding in the HP side chain model on extended*

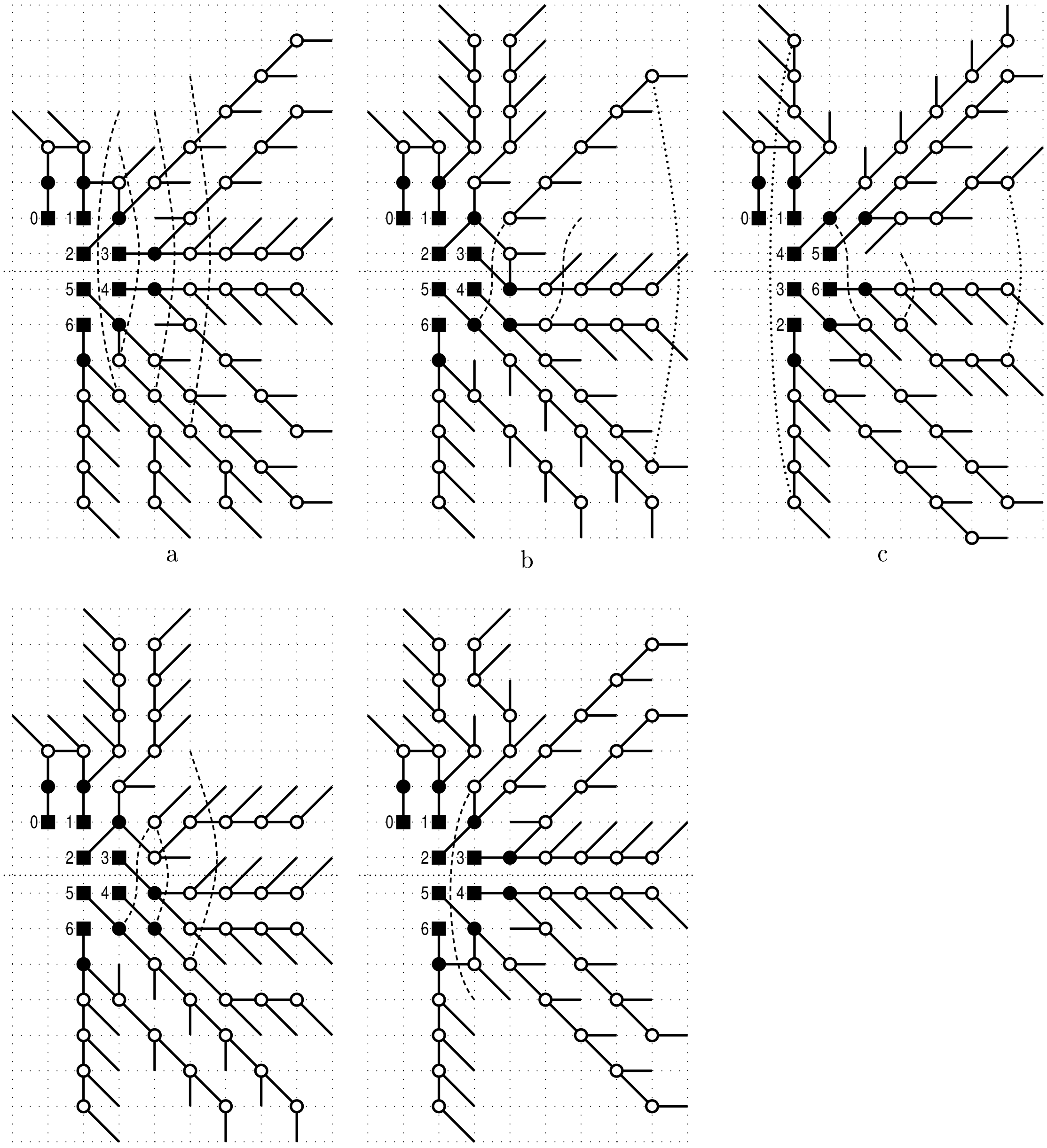

Figure 9: Folding of a 6-Fragment with a 2-Block at Position 1 and a 1-Block (Case 3.3)

cubic lattices for perfect HP-sequences with an approximation ratio of at least 37/42 *(*≈88,1%*). Moreover, this folding can be computed in linear time.*

It is possible to extend this embedding for nearly perfect HP-sequences as follows. We interpret the HP-sequence as a circle instead of a linear chain. Then it is possible

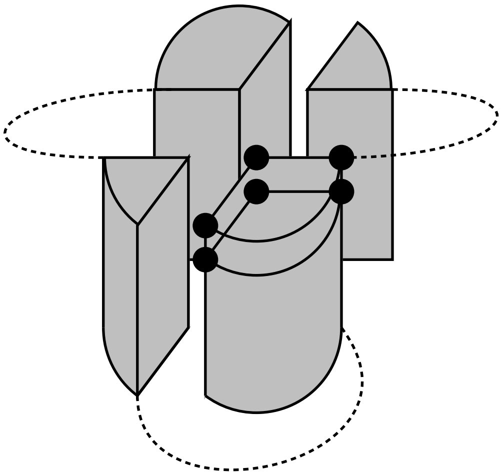

Figure 10: Final Composition of the Four Subfoldings

to position the imperfect 6-fragment at the top of one of the four poles. Using the freedom in the third dimension, it is possible to fold this fragment such that the six hydrophobic residues form two triangles as in the other foldings of 6-fragments.

Theorem 5 *Algorithm B constructs a folding in the HP side chain model on extended cubic lattices for nearly perfect HP-sequences with an approximation ratio of at least* $37/42$ *(*$\approx 88,1\%$*). Moreover, this folding can be computed in linear time.*

An inspection of the protein data base SWISS-PROT [19] shows that more than 97.5% of all stored proteins have a perfect 6-decomposition and more than 99.5% have a nearly perfect 6-decomposition. Thus, algorithm B is applicable to nearly all natural proteins.

In our analysis, we have marked the amino acids A, C, F, I, L, M, V, W, and Y (i.e., Ala, Cys, Phe, Ile, Leu, Met, Val, Trp, and Tyr) as hydrophobic and all other amino acids as polar. This classification follows Sun et al. [17] and is a conservative classification in the sense that other classifications mark more amino acids as hydrophobic. Obviously, the more amino acids are marked as hydrophobic the more proteins have a (nearly) perfect HP-sequence. The detailed analysis of amino acids in

i	0	1	2	3	4	5	>5	≥ 0
$N(i)$	71265	1315	194	86	33	24	46	72963
%	97.7%	1.8%	0.3%	0.1%				

Table 2: Statistics of Proteins in SWISS-PROT 36 with Optimal 6-Decompositions

SWISS-PROT 36 as of July 1998 can be found in Table 2. Here, $N(i)$ is the number of amino acids which have a optimal 6-decomposition with i imperfect 6-fragments. An *optimal* k-decomposition is a k-decomposition with a minimal number of imperfect k-fragments.

4 Conclusions

In this paper, we have presented two approximation algorithms for protein folding on extended cubic lattices. The general folding algorithm A achieves an approximation ratio of 59/70 (≈84.3%) for all HP-sequences. For a 'natural' subclass of HP-sequences, we have presented a more suitable folding algorithm B with a better approximation ratio of 37/42 ($\approx 88,1\%$).

It remains a challenging task to construct approximation algorithms with a high approximation ratio (better than 99%) for the HP (side chain) model on cubic-like lattices as well as on other lattices. Of course it is nor clear whether such algorithms exist. On the other hand, other 'natural restrictions' on the considered HP-sequences may yield better and more appropriate approximation algorithms.

It is also interesting to investigate which variants of the HP models allow a polynomial time approximation scheme. Despite the results of Nayak et al. [13], it is still possible that such schemes exist.

On the other hand, also good approximation algorithms for off-lattice models are desired. As a first approach Hart and Istrail showed in [12] how results from the face centered cubic lattice can be transformed in a natural way to the tangent sphere model, an off-lattice model, with a considerable loss of the approximation ratio. This transformation is quite simple because the face centered cubic lattice is very similar to the tangent sphere model.

References

[1] R. Agarwala, S. Batzoglou, V. Dančík, S. Decatur, M. Farach, S. Hannenhalli, S. Muthukrishnan, S. Skiena: Local Rules for Protein Folding on a Triangular Lattice and generalized Hydrophobicity in the HP Model, *Proceedings of the 8th Symposium on Discrete Algorithms*, 390–399, 1997, also in *Proceedings of the 1st Conference on Computational Molecular Biology*, 1–2, 1997.

[2] R. Backofen: Constraint Techniques for Solving the Protein Structure Prediction Problem, *Proceedins of the 4th International Conference on Principle and Practice of Constraint Programming*, 1998.

[3] B. Berger, F.T. Leighton: Protein Folding in the Hydrophobic-Hydrophilic (HP) Model is $\mathcal{NP}$-Complete, *Proceedings of the 2nd Conference on Computational Molecular Biology*, 30–39, 1998.

[4] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, M. Yannakakis: On the Complexity of Protein Folding, *Proceedings of the 30th Symposium on Theory of Computing*, 597–603, 1998, also in *Proceedings of the 2nd Conference on Computational Molecular Biology*, 61–62, 1998.

[5] S.E. Decatur: Protein Folding in the Generalized Hydrophobic-Polar Model on the Triangular Lattice, *Technical Report*, MIT-LCS-TM-559, MIT, 1996.

[6] K.A. Dill: Dominant Forces in Protein Folding, *Biochemistry*, 29(31):7133–7155, 1990.

[7] K.A. Dill, S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, H.S. Chan: Principles of Protein Folding — A Perspective From Simple Exact Models, *Prot. Sci.*, 4:561–602, 1995.

[8] A. Fraenkel: Complexity of Protein Folding, *Bull. Math. Biol.*, 55(6):1199–1210, 1993.

[9] K. Hamaguchi: *The Protein Molecule — Conformation, Stability, and Folding*, Japan Scientific Societies Press and Springer-Verlag, 1992.

[10] W.E. Hart, S. Istrail: Fast Protein Folding in the Hydrophobic-Hydrophilic Model Within Three-Eights of Optimal, *Proceedings of the 27th Symposium on Theory of Computing*, 157–167, 1995.

[11] W.E. Hart, S. Istrail: Fast Protein Folding in the Hydrophobic-Hydrophilic Model Within Three-Eights of Optimal, *J. Comp. Biol.*, 3(1):53–96, 1996.

[12] W.E. Hart, S. Istrail: Lattice and Off-Lattice Side Chain Models of Protein Folding: Linear Time Structure Prediction Better Than 86% of Optimal, *Proceedings of the 2nd Conference on Computational Molecular Biology*, 137–146, 1997.

[13] A. Nayak, A. Sinclair, U. Zwick: Spatial Codes and the Hardness of String Folding Problems, *Proceedings of the 9th Symposium on Discrete Algorithms*, 639–648, 1998.

[14] J.T. Ngo, J. Marks: Computational Complexity of a Problem in Molecular Structure Prediction, *Prot. Engng.*, 5(4):313–321, 1992.

[15] 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*, K. Merz Jr., S. LeGrand (Eds.), Birkhäuser, 1994.

[16] M. Paterson, T. Przytycka: On the Complexity of String Folding, *Proceedings of the 23rd International Colloquium on Automata, Languages, and Programming*, 658–669, 1996.

[17] S. Sun, R. Brem, H.S. Chan, K.A. Dill: Designing Amino Acid Sequences to Fold with Good Hydrophobic Cores, *Prot. Engng.* 8(12):1205–1213, 1995.

[18] R. Unger, J. Moult: Finding the Lowest Free Energy Conformation of a Protein is an $\mathcal{NP}$-Hard Problem: Proof and Implications, *Bull. Math. Biol.*, 55(6):1183–1198, 1993.

[19] SWISS-PROT Protein Sequence Data Bank: http://www.expasy.ch/sprot/, Data: ftp://www.expasy.ch/databases/swiss-prot/sprot36.dat, as of July 21, 1998.