Lattice and Off-Lattice Side Chain Models of Protein Folding: Linear Time Structure Prediction Better Than 86% of Optimal (Extended Abstract)*

William E. Hart† Sorin Istrail‡

August 9, 1996

Abstract

This paper considers the protein structure prediction problem for lattice and off-lattice protein folding models that explicitly represent side chains. Lattice models of proteins have proven extremely useful tools for reasoning about protein folding in unrestricted continuous space through analogy. This paper provides the first illustration of how rigorous algorithmic analyses of lattice models can lead to rigorous algorithmic analyses of off-lattice models. We consider two side chain models: a lattice model that generalizes the HP model (Dill 85) to explicitly represent side chains on the cubic lattice, and a new off-lattice model, the HP Tangent Spheres Side Chain model (HP-TSSC), that generalizes this model further by representing the backbone and side chains of proteins with tangent spheres. We describe algorithms for both of these models with mathematically guaranteed error bounds. In particular, we describe a linear time performance guaranteed approximation algorithm for the HP side chain model that constructs conformations whose energy is better than 86% of optimal in a face centered cubic lattice, and we demonstrate how this provides a 70% performance guarantee for the HP-TSSC model. This is the first algorithm in the literature for off-lattice protein structure prediction that has a rigorous performance guarantee. Our analysis of the HP-TSSC model builds off of the work of Dančík and Hannenhalli who have developed a 16/30 approximation algorithm for the HP model on the hexagonal close packed lattice. Further, our analysis provides a mathematical methodology for transferring performance guarantees on lattices to off-lattice models. These results partially answer the open question of Karplus et al. (1994) concerning the complexity of protein folding models that include side chains.

1 Introduction

Lattice models of proteins have proven extremely useful tools for reasoning about protein folding in unrestricted continuous space through analogy [4]. Lattice models sacrifice atomic detail to extract essential principles, make predictions, and to unify our understanding of many different properties of proteins. One of the important approximations made by lattices is the discretization of the conformational space. While this discretization precludes a completely accurate model of protein structures, it preserves important features of the problem of protein structure prediction,

*This work was supported by the Applied Mathematical Sciences program, U.S. Department of Energy, Office of Energy Research, and was performed at Sandia National Laboratories, operated for the U.S. Department of Energy under contract No. DE-AC04-94AL85000.

†Sandia National Labs, Algorithms and Discrete Mathematics Department, P. O. Box 5800, Albuquerque, NM 87185-1110; wehart@cs.sandia.gov; http://www.cs.sandia.gov/~wehart

‡scistra@cs.sandia.gov; http://www.cs.sandia.gov/~scistra

DISCLAIMER

like the difficulty of the related search problem. Consequently, methods that predict the structure of proteins for lattice models provide insight into the exact structure of proteins.

One common way to discretize the structure of proteins is to model the protein as a linear chain of beads in which each bead represents an amino acid. An example of this type of model is the hydrophobic-hydrophilic model (HP model) [9]. This model abstracts the hydrophobic interaction in protein folding by labeling the beads as hydrophobic (nonpolar) or hydrophilic (polar). Although a wide variety of methods have been proposed for predicting the structure of proteins in linear chain lattice models [4], none of these methods can guarantee that they can efficiently predict the native structure (which has the lowest free energy) for all proteins.

Ngo, Marks and Karplus [10] argue that an interesting approach to protein structure prediction is the development of performance guaranteed approximation algorithms. Approximation algorithms might be of significant practial use if they can be used to generate crude structures that are further refined with other techniques. We [6, 7] have recently described approximation algorithms for a variety of linear lattice models that have performance guarantees, including the linear HP model studied by Dill and his colleagues. In related work, Dančík and Hannenhalli [2] have demonstrated that performance guarantees of nearly 60% can be acheived for the HP model on the hexagonal close packed lattice.

This paper describes approximation algorithms for HP lattice and off-lattice protein models that explicitly represent side chains. The lattice model we analyze represents the conformation of a protein using a subclass of branched polymers called "branched combs." This model was proposed by Bromberg and Dill [1], who argue that linear lattice models fail to capture properties of protein folding such as side chain packing that affect the stability of the native protein structure. The HP side chain model that we consider treats the backbone of the protein as a linear chain of beads. Connected to each bead on the backbone is a bead that represents an amino acid, and each of these beads is labeled hydrophobic or hydrophilic. The off-lattice model generalizes the lattice model by representing the backbone and amino acids as tangent spheres.

The algorithms we describe generate structures that approximate the native folded state by creating compact, low energy structures that are near-optimal. Furthermore, these algorithms compute these structures in a number of computational steps that is linear in the length of the sequence. We describe approximation algorithms for the 2D and 3D cubic lattices as well as the face centered cubic (FCC) lattice. We also describe how any performance guaranteed algorithm for the FCC lattice can be used to provide performance

2 Preliminaries

2.1 The HP Side Chain Model

The protein folding model analyzed in this paper is a hydrophilic-hydrophobic model (HP model). HP models abstract the hydrophobic interaction process in protein folding by reducing a protein to a heteropolymer that represents a predetermined pattern of hydrophobicity in the protein; nonpolar amino acids are classified as hydrophobic and polar amino acids are classified as hydrophilic. A sequence is $s \in \{0,1\}^+$, where 1 represents a hydrophobic amino acid and 0 represents a hydrophilic amino acids. A HP model on 2D and 3D cubic lattices was proposed by Dill [3]. In this model, the protein is represented by a self-avoiding path on the cubic lattice, where each vertex on the path represents an amino acid. This is one of the most studied lattice models, and despite its simplicity the model is powerful enough to capture a variety of properties of actual proteins [4].

We consider a HP model that uses the model studied by Bromberg and Dill [1] to explicitly

represent side chains. In this model, a *conformation* C of a protein sequence s in a lattice L is an embedding of a catepillar graph where vertices are mapped one-to-one to lattice points, and protein bonds are mapped to the corresponding lattice edges (see Figure 1a). The legs of the catepillar graph represent amino acids, and they are labeled either hydrophobic or hydrophilic. The spine of the graph is labeled as the backbone of the protein. The *energy* of a conformation of the protein sequence s in L is defined as the sum of the energies of the hydrophobic-hydrophobic contacts, each of which contributes -1 to the total energy. A contact is defined as an edge between two amino acids in the embedded catepillar graph.

2.2 The HP Tangent Sphere Models

We introduce new off-lattice models that provide an off-lattice analogue to the HP model and the HP side chain model. In these models, the graph that represents the protein is is transformed to a set of tangent spheres of equal radius. Every vertex in the graph is replaced by a sphere, and edges in the graph are translated to constraints that force spheres to be tangent in a conformation (see Figure 1b). Spheres are labeled hydrophobic or hydrophilic, and contact between hydrophobic amino acids is when the spheres for these amino acids are in contact.

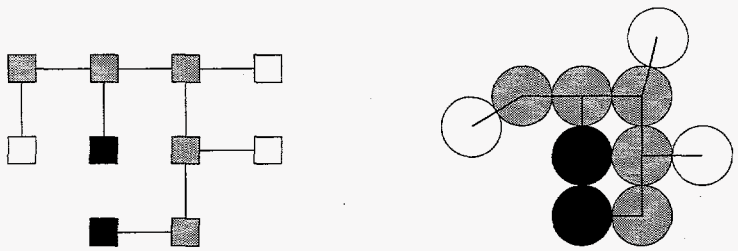


Figure 1: Illustration of conformations in (a) the HP side chain model (on a cubic lattice) and (b) the HP tangent spheres side chain model (black lines represent connections between spheres).

2.3 Computational Complexity

According to the Thermodynamic Hypothesis the *native conformation* of a protein is the conformation with the minimum energy among the set of all conformations. Thus we algorithmicaly formulate the problem of predicting the native conformation as finding an efficient algorithm that computes the native conformation of a sequence s in a lattice L. A protein folding algorithm is *efficient* if for every sequence it determines the native conformation in polynomially many steps in the length of the sequence.

It is unknown whether any well studied protein structure prediction problem can be solved efficiently, including the HP side chain model. Hart and Istrail [8] have recently shown that a broad class of protein structure prediction problems are NP-complete, which means that they are practically intractable [5]. Although they consider is a broad class of side chain models, their results are not immediately applicable to the HP side chain model.

This paper presents performance guaranteed approximation algorithms for the HP side chain model. Two standard types of performance guarantees are [5]: the *absolute performance ratio* and the *asymptotic performance ratio*. Let $\mathcal{Z}_L(s)$ be the energy of the conformation generated for

protein instance s on lattice L with by algorithm $\mathcal{Z}_L$, and let $OPT_L(s)$ be the energy of the optimal conformation of s on L. Recall that both $\mathcal{Z}_L(s)$ and $OPT_L(s)$ are nonpositive integers for every s. The *absolute performance ratio* $R(\mathcal{Z}_L)$ of algorithm $\mathcal{Z}_L$ is given by

$$R(\mathcal{Z}_L) = \sup\{r \geq 1 \mid \forall s, R_{\mathcal{Z}_L}(s) \geq r\},$$

where $R_{\mathcal{Z}_L}(s) = \mathcal{Z}_L(s)/OPT_L(s)$. Given $N \in \mathbf{Z}$, let $S_N^L = \{s \mid OPT_L(s) \leq N\}$, and let $R_{\mathcal{Z}_L}^N = \inf\{R_{\mathcal{Z}_L}(s) \mid s \in S_N^L\}$. The *asymptotic performance ratio* $R^\infty(\mathcal{Z}_L)$ is given by

$$R^\infty(\mathcal{Z}_L) = \sup\{r \mid R_{\mathcal{Z}_L}^N \geq r, N \in \mathbf{Z}\} = \sup_{N \in \mathbf{Z}} \inf_{s \in S_N^L} R_{\mathcal{Z}_L}(s).$$

If $R(\mathcal{Z}_L) = \tau$ for a fixed constant τ, then the value of solutions generated by algorithm $\mathcal{Z}_L$ are within a factor of τ of the optimum. If $R^\infty(\mathcal{Z}_L) = \tau$, then as $\mathcal{Z}_L$ is applied to larger protein instances, the value of solutions generated by $\mathcal{Z}_L$ approaches a factor of τ of the optimum. Here, "large" protein instances have low conformational energy at their native state, which may be independent of their length. Since $\mathcal{Z}_L(s) \leq 0$ and $OPT_L(s) \leq 0$, both of these ratios are scaled between 0 and 1 such that a ratio closer to 1 indicates better performance.

3 The HP Side Chain Model on Cubic Lattices

This section describes performance guaranteed approximation algorithms for the HP side chain model on the 2D and 3D cubic lattices. We begin by describing bounds on the optimum for these models. Following Hart and Istrail [6], we decompose a protein sequence into a series of x- and y-blocks, $x_1y_1x_2 \ldots x_ny_n$. Within each block, hydrophobic amino acids are seperated by an odd number of hydrophilic amino acids, and between blocks there are an even number of hydrophilic amino acids. For a protein sequence, $N_x(s)$ is the number of hydrophobics in x-blocks and N_y is the number of hydrophilics in y-blocks. We say that $X = N_x(s)$ and $Y = N_y(s)$ and assume that the labeling of blocks guarantees that $X \leq Y$.

Let $OPT_{2D}(s)$ be the value of the optimal conformation of s in the 2D model, and let $OPT_{3D}(s)$ be the value of the optimal conformation of s in the 3D model. In the 2D model, every 1 in each x-block can be a topological neighbor of at most three other 1s. Thus the optimal energy is at most $OPT_{2D}(s) \geq -3X$. In the 3D model, every 1 in each x-block can be a topological neighbor of at most five other 1s. Thus the optimal energy is at most $OPT_{3D}(s) \geq -5X$.

3.1 Approximation Algorithms

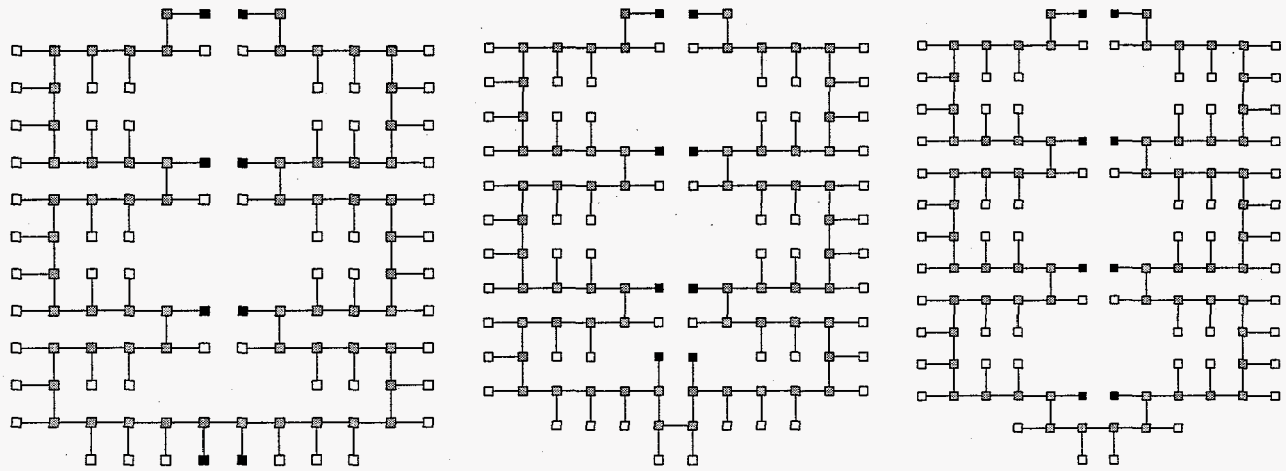
We begin by describing Algorithm $\mathcal{A}$, an approximation algorithm for the 2D HP side chain model. Algorithm $\mathcal{A}$ selects a single folding point (turning point) that divides a protein instance into subsequences B' and B'', such that $N_y(B')$ is balanced with $N_x(B'')$. The conformation for these two halves of the protein sequence are constructed such that the y hydrophobics in B' and the x hydrophobics in B'' are configured face-to-face to form a hydrophobic core.

The folding point is selected using "Subroutine 1" from Hart and Istrail [6]. Subroutine 1 selects a folding point that balances the hydrophobicity between the x-blocks and y-blocks on each half of the folding point. The following lemma describes the key property of the folding point that is selected.

Lemma 1 ([6], Lemma 1) The folding point selected by Subroutine 1 partitions a protein instance s into two subsequences B' and B'' such that either

$$N_y(B') \geq \lceil (Y+1)/2 \rceil \quad \text{and} \quad N_x(B'') \geq \lceil X/2 \rceil \quad \text{or} \quad N_y(B') \geq \lceil Y/2 \rceil \quad \text{and} \quad N_x(B'') \geq \lceil (X+1)/2 \rceil .$$

Figure 2 illustrates the conformations generated by Algorithm $\mathcal{A}$ for different types of folding points. Decomposition into x- and y-blocks requires a single pass through the protein instance. Subroutine 1 requires a single pass through the sequence of blocks, which is no longer than the length of the protein instance. The construction of the structures for B' and B'' also requires linear time. Thus the computation required by Algorithm $\mathcal{A}$ is linear. The performance of Algorithm $\mathcal{A}$ can be bounded as follows.



(a) (b) (c)

Figure 2: Illustration of the different folding points used for different block separators z_i at the folding point, for (a) $l(z_i) = 0$, (b) $l(z_i) = 2$, and (c) $l(z_i) \geq 4$. Gray blocks represent the backbone, white blocks represent hydrophilic amino acids and black blocks represent hydrophobic amino acids.

Lemma 2 $\mathcal{A}(s) \leq -\lceil X/4 \rceil$.

The following proposition presents the asymptotic and absolute performance guarantees for Algorithm $\mathcal{A}$.

Proposition 1 $1/6 \geq R^{\infty}(\mathcal{A}) \geq R(\mathcal{A}) \geq 1/12$.

We now describe Algorithm $\mathcal{B}$, a performance guaranteed approximation method for the 3D HP model with side chains. Algorithm $\mathcal{B}$ selects a single folding point that divides the protein instance into two subsequences B' and B'', such that $N_y(B')$ is balanced with $N_x(B'')$. The conformation generated by Algorithm $\mathcal{B}$ places the y hydrophobics in B' and the x hydrophobics in B'' to form a hydrophobic core that is a solid block of hydrophobic amino acids with dimension 2x2xk (for some k). Each edge of this block is formed by interleaving the hydrophobics from B' and B'' This interleaving allows each hydrophobic amino acids to form contacts with four other hydrophobic amino acids.

Figure 3 illustrates how the structures for B' and B'' are interleaved to form a single column of the hydrophobic core, including an illustration how the folding point is formed. Figures 4a and 4b provide high level illustrations of the structures used for B' and B''. Figure 5a illustrates

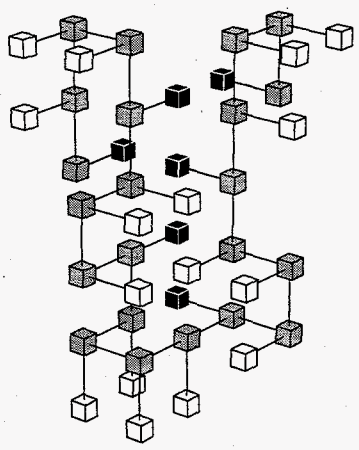


Figure 3: Illustration of how a single column of hydrophobics is formed by Algorithm $\mathcal{B}$. This figure also illustrates the conformation of the folding point.

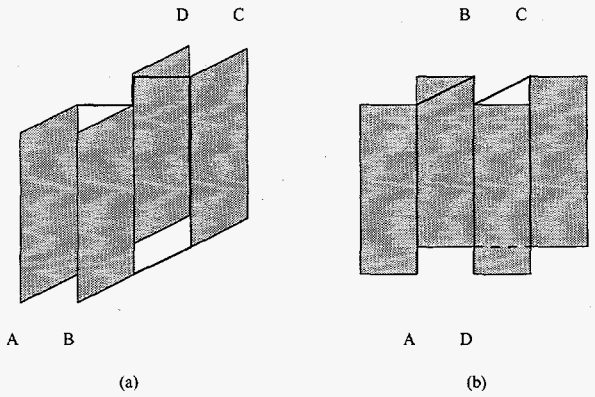


Figure 4: A graphic illustration of the general structure of the subsequences B' and B'' in (a) and (b) respectively. The gray planes illustrate the position of the backbone of the loops of nonhydrophobics. The labels A, B, C and D indicate the order of the labels, starting from the folding point between the A planes of B' and B''.

(a) (b)

Figure 5: Illustration of the entire conformation generated by Algorithm $\mathcal{B}$: (a) a view from the side and (b) a view from the top highlighting the hydrophobic core.

the application of Algorithm $\mathcal{B}$ to a protein sequence, and Figure 5b provides an end-view of this conformation that illustrates the core formed by Algorithm $\mathcal{B}$.

Each step of Algorithm $\mathcal{B}$ is linear, so Algorithm $\mathcal{B}$ requires linear time. The performance of Algorithm $\mathcal{B}$ can be bounded as follows.

Lemma 3 Let $\bar{X} = \lceil X/2 \rceil$. If $\bar{X} \geq 8$ then

$$\mathcal{B}(s) \leq -4\bar{X} + 28.$$

The following proposition presents the asymptotic and absolute performance ratios for Algorithm $\mathcal{B}$.

Proposition 2 $R_{\mathcal{B}}^{\infty} = 4/10$ and $4/10 \geq R_{\mathcal{B}} \geq 1/12$.

3.2 Related Results

Embedded Algorithms for the 3D HP side chain model Conformations for the 2D HP side chain model can be trivially embedded in 3D to generate conformations for the 3D HP side chain model. Similarly, a conformation from the 2D HP model can be used to construct a conformation in the 3D HP side chain model as follows: (1) embed the conformation on any 2D plane, (2) create side chains for each monomer, all of which are placed on the same adjacent planes, and (3) label the side chains with the hydrophobicities of their corresponding backbone monomers, and unlabel the backbone monomers. It is possible to show that performance guaranteed approximation algorithms for the 2D HP model and the 2D HP side chain model can be used to provide performance guarantees for the 3D HP side chain model.

Variable Length Side Chains A natural extension of the side chain model that we have considered is to include notions of volume into the side chain formulation. One way of doing this would be to model the volume of a side chain by varying the length of the legs of the catepillar graph. All of the vertices in the legs are labeled hydrophobic or hydrophilic, but not necessarily uniformly within a given leg. If we assume that this chain has a bounded length, β, then a simple modification of Algorithm $\mathcal{A}$ leads to a performance guarantee in terms of $1/\beta$. The blocks in this modified algorithm are based on the amino acid vertices adjacent to the protien's backbone. The structures for B' and B'' are expanded to allow side chains of up to length β to fit into each "zero loop" to either side of the hydrophobic core, and the side chains within the core turn immediately to form hydrophobic contacts. The analysis of this algorithm gives a performance guarantee of $\frac{1}{12\beta}$. Following arguments similar to those mentioned in the previous paragraph, this algorithm also provides a performance guarantee for the 3D HP side chain model.

4 The HP Side Chain Model on the FCC Lattice

We now describe Algorithm $\mathcal{C}$, a performance guaranteed approximation method for the HP model with side chains on the face centered cubic lattice. Algorithm $\mathcal{C}$ builds upon the analysis of Dancik and Hannenhalli [2] that describes an approximation for the HP model on the FCC lattice. Figure 6 illustrates the packing of vertices in a FCC lattice. The center of each sphere represents the location of a single vertex, and contacts between spheres represent edges between vertices. The gray spheres illustrate a *layer* of the FCC lattice, which is composed of two adjacent horizontal planes of vertices. The bold spheres illustrate a vertical *column* of this lattice.

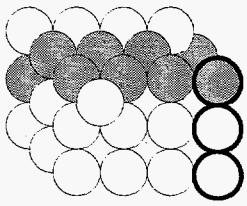


Figure 6: Illustration of the general structure of the FCC lattice, highlighting a layer (in gray) and a column (bolded spheres).

Let $N(s)$ equal the number of hydrophobics in a sequence s. Algorithm $\mathcal{C}$ divides s into eight subsequences such that each subsequence contains approximately $N(s)/8$ hydrophobics. Each subsequence B_i is configured such that all of the hydrophobics in B_i are placed together in a single column. Consecutive hydrophobics in B_i are in contact within this column. These eight columns are configured to form a 2x4 solid hydrophobic core that contains no hydrophilics (see Figure 7).

To form these columns of hydrophobics, we configure the loops of hydrophobics such that they never intersect. Figure 7a illustrates the structure of these loops for half of the conformation (the other half can be constructed symmetrically). Note that the structure of the loops differs for each of the four columns; Figure 7b illustrates the structure of the bottom column in Figure 7a for hydrophilic loops of all lengths (the last structure can be extended for loops of length six or more). The structure shown in Figure 7a illustrates how a single layer of the columns is configured. Each column is constructed by forming loops of hydrophilics that lie within a single layer. The hydrophilic loops for subsequent hydrophobics are disjoint because each hydrophobic along a column utilizes a disjoint layer to form its loop.

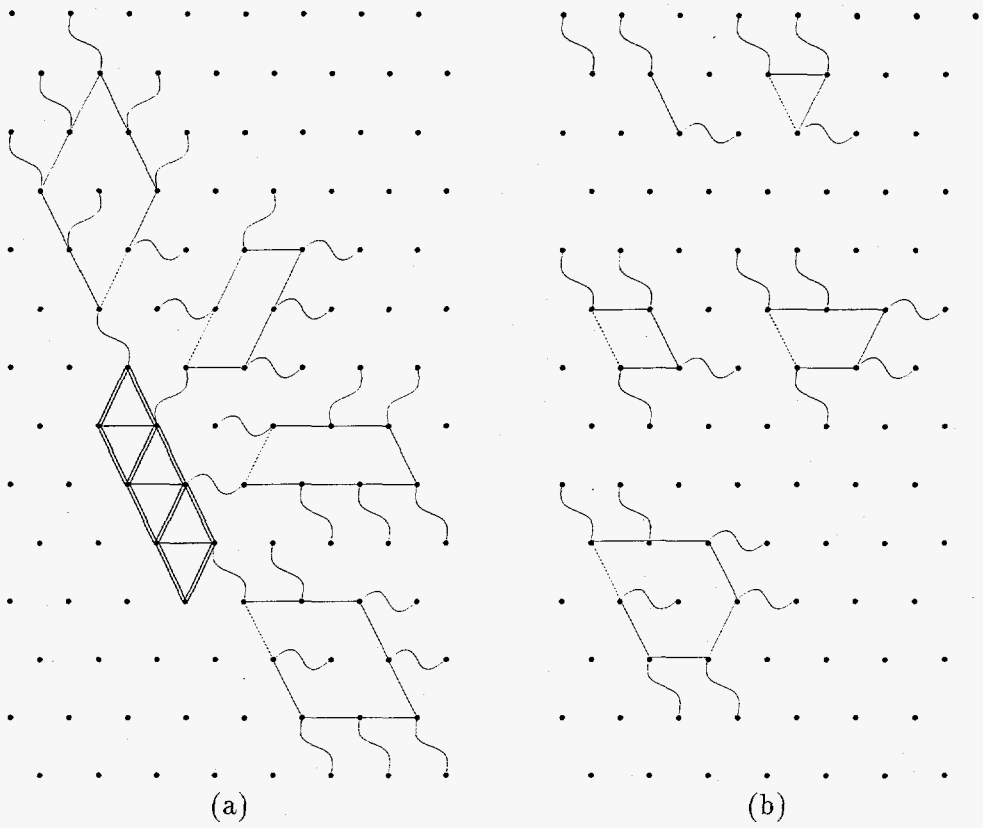


Figure 7: Illustration of (a) the general structure of Algorithm $\mathcal{C}$ for a single layer of the FCC lattice and (b) the structure of all loops of the bottom column. The points on this figure represent columns on the lattice. The loops of hydrophilics for four of the columns are illustrated in (a); the other four columns have a complementary structure. Solid lines represent the path of the backbone within a single layer, and dashed lines represent the path of the backbone between adjacent. The curved lines represent the positions of the side chains. The interactions between the eight hydrophobic columns are highlighted with either one or two dark lines, indicating the number of contacts each hydrophobic makes between a pair of columns.

The construction of the conformations for each column can proceed sequentially, so Algorithm $\mathcal{C}$ requires linear time. Note that unlike the approximation algorithms for the cubic lattice, Algorithm $\mathcal{C}$ does not require a global calculation of the folding point. The only global information needed for this algorithm is the computation of the total number of hydrophobics in the sequence. The following lemma describes the performance guarantee for Algorithm $\mathcal{C}$.

Lemma 4 $\mathcal{C}(s) \leq -31N(s)/8 + 69$.

We consider a bound on the value of $OPT(s)$. A trivial bound of $OPT(s) \geq -11N(s)/2$ is easy to establish by noting that each hydrophobic side chain can make at most 11 hydrophobic contacts, each of which must be shared. We can improve this bound by observing that there are four contact points with a side chain that also form contacts with the backbone at the side chain. The implies that each hydrophobic side chain forces four *conflicts* [2]. If a contact point is empty or contains a backbone or hydrophilic, then the current side chain does not make 11 contacts. If the

contact point contains a hydrophobic then that hydrophobic side chain cannot make 11 contacts. This observation can be used to prove the following lemma.

Lemma 5 $OPT(s) \geq -9N(s)/2$.

Combining Lemmas 4 and 5, we get the following performance guarantee for Algorithm $\mathcal{C}$.

Proposition 3 $R_{\mathcal{C}}^{\infty} \geq 31/36$.

5 Algorithmic Performance for Off-Lattice Models

The class of tangent spheres models (with or without chains) has the property that it can be analysed rigorously by transferring algorithmic analyses from various lattice HP-models to the off-lattice setting. In this section we focus on the tangent spheres model with side chains and show how a conformation created by Algorithm $\mathcal{C}$ on the FCC lattice provides a performance guarantee for this model off-lattice. The linear chain tangent spheres model can be similarly analysed. Using the hexagonal close packed lattice algorithm of Dančík and Hannenhalli [2], one can prove that it has at least 46.7% of optimal off-lattice performance.

To analyze the performance of the off-lattice tangent spheres side chain model, we begin by deriving lower bounds on the number of possible contacts that each hydrophobic side chain can make. It is well-know that for a set of identical spheres in 3D the maximum number of spheres that can be tangent to a single fixed sphere is 12. This is the so called the *kissing number*. From this we can conclude that a hydrophobic side chain can be tangent to only 11 other hydrophobic side chain, since one position is taken by the backbone sphere connected to it. As contacts are binary (between two spheres), each side chain can contribut at most 11/2 contacts by reasoning abstractly in the worst case.

Our tangent spheres side chain model generalizes the HP model in the sense that for any lattice a conformation in that lattice represents a possible off-lattice conformation. To provide a performance guarantee for the off-lattice, we apply Algorithm $\mathcal{C}$ to generate a conformation on the FCC lattice, which is guaranteed to have and energy of no more than $-31N(s)/8 + 69$. Using the lower bound of $-11N(s)/2$ on the value of the optimum, we can show that Algorithm $\mathcal{C}$ provides an asymptotic performance ratio of $31/44 > 70\%$.

Our analysis of the lower bound is actually quite optimistic. We conjecture that a stronger analysis can improve the performance guarantee to over 77% of optimal. This conjecture is based on our belief that if an amino acid has 11 contacts then there is at least one contact that is sufficiently close to the backbone of the side chain to form a "conflict" that prevents that sphere from making 11 contacts itself. If this is true then each side chain contributes at most 5 contacts, thereby giving the stated performance guarantee. Furthermore, we suspect that the notion of a conflict can be extended in this fashion to provide even stronger performance guarantees.

Acknowledgements

Our thanks to Sarina Bromberg and Ken Dill for discussions that inspired the side chain models that we have analyzed.

References

[1] S. Bromberg and K. A. Dill. Side chain entropy and packing in proteins. *Prot. Sci.*, pages 997–1009, 1994.

[2] V. Dančík and S. Hannenhalli. Protein folding on a triangular mesh, May 1996. Unpublished research.

[3] K. A. Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 24:1501, 1985.

[4] 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.

[5] M. R. Garey and D. S. Johnson. *Computers and Intractability - A guide to the theory of NP-completeness.* W.H. Freeman and Co., 1979.

[6] W. E. Hart and S. Istrail. Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal. To appear in *Journal of Computational Biology*, Spring 1996. Extended abstract in *Proc. of 27th Annual ACM Symposium on Theory of Computation*, May 1995.

[7] W. E. Hart and S. Istrail. Invariant patterns in crystal lattices: Implications for protein folding algorithms (extended abstract). In *Proc. 7th Annual Symp. on Combinatorial Pattern Matching*, 1996. (to appear).

[8] W. E. Hart and S. Istrail. Robust proofs of NP-hardness for protein folding: General lattices and energy potentials. 1996. (in preparation).

[9] K. F. Lau and K. A. Dill. A lattice statistical mechanics model of the conformation and sequence spaces of proteins. *Macromolecules*, 22:3986–3997, 1989.

[10] J. T. Ngo, J. Marks, and M. Karplus. Computational complexity, protein structure prediction, and the Levinthal paradox. In K. Merz, Jr. and S. Le Grand, editors, *The Protein Folding Problem and Tertiary Structure Prediction*, chapter 14, pages 435–508. Birkhauser, Boston, MA, 1994.

Appendix to "Lattice and Off-Lattice Side Chain Models of Protein Folding: Linear Time Structure Prediction Better Than 86% of Optimal (Extended Abstract)"

William E. Hart and Sorin Istrail
{wehart,scistra}@cs.sandia.gov

A Protein Sequence Structure on Cubic Lattices

This section summarizes key definitions concerning the structure of protein instances from Hart and Istrail [6]. Let $s = s_1, \ldots, s_m$ be a protein instance, $s_i \in \{0, 1\}$, where 1's correspond to hydrophobics and 0's correspond to hydrophilics. Let $l(s)$ equal the length of the sequence s. Let $M_{max}(s)$ equal the length of the longest sequence of zeros in s, and let $M_{min}(s)$ equal the length of the shortest sequence of zeros in s.

An instance s can be decomposed into a sequence of *blocks*. A block b_i has the form $b_i = 1$ or $b_i = 1Z_{i_1}1 \ldots Z_{i_h}1$, where the Z_{i_j} are odd-length sequences of 0's and $h \geq 1$. A *block separator* z_i is a sequence of 0's that separates two consecutive blocks, where $l(z_i) \geq 0$ and $l(z_i)$ is even for $i = 1, \ldots, h-1$. Thus s is decomposed into $z_0 b_1 z_1 \ldots b_h z_h$. Since $l(z_i) \geq 0$, this decomposition treats consecutive 1's as a sequence of blocks separated by zero-length block separators. Let $N(b_i)$ equal the number 1's in b_i. Thus the sequence

$$0\underbrace{10101}_{b_1}\underbrace{1}_{b_2}\underbrace{1}_{b_3}\underbrace{10101}_{b_4}0000\underbrace{1010101}_{b_5}$$

can be represented as $l(z) = (1, 0, 0, 0, 4, 0)$ and $N(b) = (3, 1, 1, 3, 4)$.

Note that two 1's can be endpoints of a contact edge only if there is an even number of elements between them [6]. It follows from our definition of blocks that two 1's within a block cannot be in contact. Further, any pair of 1's take from blocks b_k and b_j may be in contact only when $|k - j|$ is odd.

Since 1's from a block can only be in contact of 1's from every other block, it is useful to divide blocks into two categories: x-blocks and y-blocks. For example, let $x_i = b_{2i}$ and let $y_i = b_{2i-1}$. This makes it clear that 1's from an x-block can only be in contact with 1's from an y-block. Let B_x and B_y be the number of x-blocks and y-blocks respectively. Further, let $X = X(s) = \sum_{i=1}^{B_x} N(x_i)$ and $Y = Y(s) = \sum_{i=1}^{B_y} N(y_i)$. We assume that the division into x- and y-blocks is such that $X \leq Y$. For example, the sequence

$$0\underbrace{10101}_{y_0}\underbrace{1}_{x_0}\underbrace{1}_{y_1}\underbrace{10101}_{x_1}0000\underbrace{1010101}_{y_2}$$

can be represented as $z_0 y_0 z_1 x_0 z_2 y_1 z_3 x_1 z_4 y_2 z_5$, where $l(z) = (1, 0, 0, 0, 4, 0)$, $N(x) = (1, 3)$, and $N(y) = (3, 1, 4)$.

B Proofs of Lemmas, Propositions and Theorem

B.1 PROOF OF LEMMA 2

Proof. From Lemma 1 we know that $N_y(B') \geq \lceil X(s)/2 \rceil$ and $N_x(B'') \geq \lceil X(s)/2 \rceil$. Consequently the number of contact edges is at least

$$\left\lceil \frac{\lceil X(s)/2 \rceil}{2} \right\rceil \geq \left\lceil \frac{X(s)}{4} \right\rceil.$$

The folding point does not eliminate any of these contact edges, so the final energy is at least $-\lceil X(s)/4 \rceil$. ∎

B.2 PROOF OF PROPOSITION 1

Proof. The middle inequality follows from the definitions of $R_{\mathcal{A}}$ and $R^{\infty}(\mathcal{A})$. We know from

Lemma 2 that $\mathcal{A}(s) \leq -\lceil X(s)/4 \rceil$. Now $OPT(s) \leq -3X(s)$, so

$$R_{\mathcal{A}}(s) = \frac{\mathcal{A}(s)}{OPT(s)} \geq \frac{-\left\lceil \frac{X(s)}{4} \right\rceil}{-3X(s)} \geq \frac{-\frac{X(s)}{4}}{-3X(s)} = 1/12. \tag{1}$$

To show that $R^{\infty}(\mathcal{A}) \leq 1/12$, consider instances s^k which are represented as

$$\begin{aligned} L(z) &= (0,4,4,0) \\ N(b) &= (k,6k,k) \\ M_{max}(b) = M_{min}(b) &= (3,1,3) \end{aligned}$$

Figure 8a illustrates how a protein instance s^k can be folded to get the optimal energy of $-6k$. Now $\mathcal{A}(s^k) = -k$ (see Figure 8b). Given N, $s^{\lceil -N/6 \rceil} \in S_N$. Thus $R^{\infty}(\mathcal{A}) \leq \lim_{N\to\infty} R^N(\mathcal{A}) = \frac{-\lceil -N/6 \rceil}{N} = 1/6$.

■

B.3 PROOF OF LEMMA 3

Proof. Algorithm $\mathcal{B}$ selects a folding point using Subroutine 1. We know from Lemma 1 that the folding point selected by Algorithm $\mathcal{B}$ splits the protein instance into two sequences that have at least $\lceil X/2 \rceil$ hydrophobics that can be used to form the hydrophobic core. Consequently, the structure generated by Algorithm $\mathcal{B}$ contains a 2x2xK block of hydrophobics, where $K \geq 2\lceil (\bar{X} - 4)/4 \rceil$. The -4 term accounts for contacts that can be lost at the folding point and turns. Thus the energy is

$$-16K + 4 \leq -16 \left\lceil \frac{\bar{X} - 4}{4} \right\rceil + 4 \leq -4\bar{X} + 28.$$

■

B.4 PROOF OF PROPOSITION 2

Proof. To show the lower bound on $R^{\infty}_{\mathcal{B}}$, we may assume that $\lceil X/2 \rceil \geq 8$ (since we are considering the asymptotic performance ratio). From Lemma 3 we have $\mathcal{B}(s) \leq -4\lceil X/2 \rceil + 28$. Thus

$$R_{\mathcal{B}}(s) \geq \frac{\mathcal{B}(s)}{OPT(s)} \geq \frac{-4\lceil X/2 \rceil + 28}{-5X} \geq \frac{2X - 28}{5X}.$$

For $s \in S_N$, $-5X \leq OPT(s) \leq N$, so $X \geq -N/5$. Since $\frac{2X-28}{5X}$ is monotonically increasing for $X \geq 0$, it follows that

$$R_{\mathcal{B}}(s) \geq \frac{-2N/5 - 28}{-N} = \frac{2N + 140}{5N}.$$

Thus

$$R^N_{\mathcal{B}} \geq (2N + 140)/(5N)$$

and

$$R^{\infty}_{\mathcal{B}} = \sup\{r \mid R^N_{\mathcal{B}} \geq r, N \in \mathbf{Z}\} \geq \lim_{N\to\infty} (2N + 140)/(5N) = 2/5.$$

To show the upper bound for $R^{\infty}_{\mathcal{B}}$, consider instances s^k which are represented as

$$\begin{aligned} L(z) &= (0,4,4,0) \\ N(b) &= (8k+4, 48k+24, 8k+4) \\ M(b) = m(b) &= (3,1,3), \end{aligned}$$

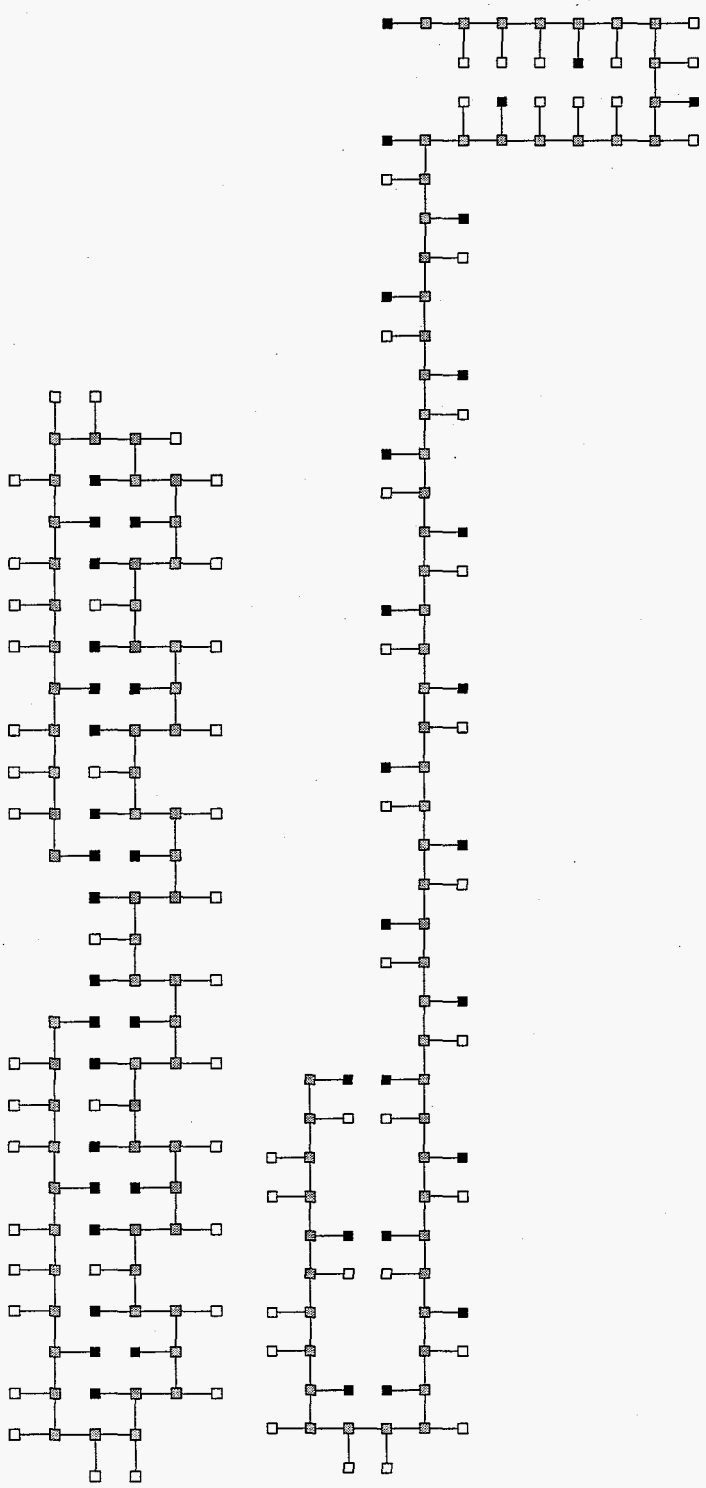


Figure 8: Conformations of s^3: (a) optimal conformation, and (b) conformation generated by Algorithm $\mathcal{A}$.

$k \geq 12$. An instance s^k can be folded to get the optimal energy of $-80k-40$, while the conformation generated by Algorithm $\mathcal{B}$ has an energy of $\mathcal{B}(s^k) \geq -32k$. Given N, $s^{\lceil -(N+40)/80 \rceil} \in S_N$. Thus $R_{\mathcal{B}}^N \leq -32 \lceil -(N+40)/80 \rceil / N$. It follows that $R_{\mathcal{B}}^{\infty} \leq \lim_{N\to\infty} R_{\mathcal{B}}^N = 2/5$.

Now $R_{\mathcal{B}} \geq 1/12$ if $OPT(s) \geq 12\mathcal{B}(s)$ for all protein instances s. If $\lceil X/2 \rceil < 8$, then Algorithm $\mathcal{B}$ applies Algorithm $\mathcal{A}$, so $OPT(s) \geq 12\mathcal{B}(s)$. If $\lceil X/2 \rceil \geq 8$ then from Lemma 3, we have $\mathcal{B}(s) \leq -4 \lceil X/2 \rceil + 28$. Thus

$$12\mathcal{B}(s) \leq -48 \lceil X/2 \rceil + 336 \leq OPT(s).$$

Consequently $R_{\mathcal{B}} \geq 1/12$.

Finally, the upper bound on $R_{\mathcal{B}}(s)$ follows from the fact that $R_{\mathcal{B}}(s) \leq R_{\mathcal{B}}^{\infty}(s)$. ■

B.5 PROOF OF LEMMA 4

Proof. Let $K = \lfloor (N(s)-6)/8 \rfloor$, which represents the height of each column of hydrophobics. The -6 term accounts for the fact that a single hydrophobic might need to be sacrificed to connect the columns on each side of the core. Now within each column there are $K-1$ hydrophobic contacts. There are 10 interactions between columns in the core that contribute $2K-1$ contacts and there are 3 interactions between columns that contribute K contacts. Thus we have

$$\begin{aligned} \mathcal{C}(s) \leq -8(K-1) - 10(2K-1) - 3K &= -31K + 18 \\ &= -31 \lfloor (N(s)-6)/8 \rfloor + 18 \\ &\geq -31N/8 + 69. \end{aligned}$$

■

B.6 PROOF OF LEMMA 5

Consider a hydrophobic side chain. The hydrophobic on this side chain can make at most 11 hydrophobic contacts. Four of these contact points form conflicts. Each of these conflicts removes a single hydrophobic contact from the set of all possible hydrophobic contacts. Since a conflict can be "shared" between two hydrophobic side chains, this means that $OPT(s) \geq -(11-4/2)/2 = -9/2$.

B.7 PROOF OF PROPOSITION 3

Proof. Lemma 4 we have $\mathcal{C}(s) \leq -31N(s)/8 + 69$. Thus

$$R_{\mathcal{C}}(s) \geq \frac{\mathcal{C}(s)}{OPT(s)} \geq \frac{-31N(s)/8 + 69}{-9N(s)/2} = \frac{31N(s) - 522}{36N(s)}.$$

For $s \in S_N$, $N \geq OPT(s) \geq -9N(s)/2$, so $N(s) \geq -2N/9$. Since $\frac{31N(s)-522}{36(s)}$ is monotonically increasing for $N(s) \geq 0$, it follows that

$$R_{\mathcal{C}}(s) \geq \frac{31(-2N/9) - 522}{36(-2N/9)} = \frac{31N + 2484}{36N}$$

Thus

$$R_{\mathcal{B}}^N \geq (31N + 2484)/(36N)$$

and

$$R_{\mathcal{B}}^{\infty} = \sup\{r \mid R_{\mathcal{B}}^N \geq r, N \in \mathbf{Z}\} \geq \lim_{N\to\infty} (31N + 2484)/(36N) = 31/36.$$

■

C Embedded Conformations in the 3D HP Side Chains Model

In this section we show how performance guaranteed approximation algorithms for the 2D HP model (which represent amino acids as bead along a chain) and the 2D HP side chain model can be used to provide performance guarantees for the 3D HP side chain model. Conformations for the 2D HP side chain model can be trivially embedded in 3D to generate conformations for the 3D HP side chain model. To generate conformations in the 3D HP side chain model using conformations from the 2D HP model, simply

1. Embed the conformation on any 2D plane
2. Create side chains for each monomer, all of which are placed on the same adjacent planes.
3. Label the side chains with the hydrophobicities of their corresponding backbone monomers, and unlabel the backbone monomers.

Let $\overline{OPT}_{2D}(s)$ be the value of the optimal conformation of s in the 2D HP model. The following theorem relates the values of $OPT_{3D}(s)$ to the values of both $\overline{OPT}_{2D}(s)$ and $OPT_{2D}(s)$. This theorem shows that the energy of the optimal conformation in the 3D HP side chain model is within a constant factor of the energy of the optimal conformations in the 2D HP model and 2D HP side chain model.

Theorem 1 $\min\left(OPT_{2D}(s), \overline{OPT}_{2D}(s)\right) \geq OPT_{3D}(s) \geq \max\left(20OPT_{2D}(s), 10\overline{OPT}_{2D}(s)\right)$.

Proof. We prove this result by proving the following inequalities:

$$OPT_{2D}(s) \geq OPT_{3D}(s) \geq 20OPT_{2D}(s) \tag{2}$$

and

$$\overline{OPT}_{2D}(s) \geq OPT_{3D}(s) \geq 10\overline{OPT}_{2D}(s). \tag{3}$$

The first inequality in Equation (2) is immediately obvious. Every conformation for the 2D HP side chain model is a conformation of the 3D HP side chain model. To prove the second inequality, consider Algorithm $\mathcal{A}$. From Lemma 2 we know that Algorithm $\mathcal{A}$ generates solutions with an energy of at least $-\lceil X/4 \rceil$. Consequently, $OPT_{2D}(s) \leq -\lceil X/4 \rceil$. Now $OPT_{3D}(s) \geq -5X$, so

$$20OPT_{2D}(s) \leq -20\lceil X/4 \rceil \leq -5X \leq OPT_{3D}(s).$$

The first inequality in Equation (3) follows from the fact that every conformation for the 2D HP model can be used to construct a conformation for the 3D HP side chain model with the same energy, using the method described above. Now consider "Algorithm $\mathcal{B}$" for the 2D HP model described by Hart and Istrail [6]. This algorithm is guaranteed to generate conformations with an energy of at least $-\lceil (X+1)/2 \rceil$. Consequently,

$$10\overline{OPT}_{2D}(s) \leq -10\lceil (X+1)/2 \rceil \leq -5(X+1) \leq OPT_{3D}(s).$$

■

The upper bound on $OPT_{3D}(s)$ is tight. It is not clear whether the lower bound is tight, since the proof of the lower bound uses the absolute performance guarantees for the best known approximation algorithms for the 2D HP model and 2D HP side chain model (1/4 and 1/12 respectively).

Using the bounds on $OPT_{3D}(s)$ provided by Theorem 1, we can demonstrate that performance guaranteed algorithms for the 2D HP model and the 2D HP side chain model provide performance

guarantees for the 3D HP side chain model. Let $\mathcal{Z}_{2D}$ ($\mathcal{Z}_{3D}$) refer to the application of a generic Algorithm $\mathcal{Z}$ in the 2D HP side chain model (Algorithm $\mathcal{Z}$ in the 3D HP side chain model), and let $\overline{\mathcal{Z}}_{2D}$ refer to the application of a generic Algorithm $\overline{\mathcal{Z}}$ in the 2D HP side chain model. The following two propositions provide bounds on $R_{\mathcal{Z}_{3D}}$ and $R^{\infty}_{\mathcal{Z}_{3D}}$.

Proposition 4 If $\kappa_1 \leq R_{\mathcal{Z}_{2D}} \leq \kappa_2$ and $\kappa_3 \leq R^{\infty}_{\mathcal{Z}_{2D}} \leq \kappa_4$
then $\kappa_1/20 \leq R_{\mathcal{Z}_{3D}} \leq \kappa_2$ and $\kappa_3/20 \leq R^{\infty}_{\mathcal{Z}_{3D}} \leq 3\kappa_4/5$.

Proof. Let $S^{2D}_N = \{s \mid OPT_{2D}(s) \leq N\}$ and $S^{3D}_N = \{s \mid OPT_{3D}(s) \leq N\}$.

We first prove the absolute performance guarantees for $\mathcal{Z}_{3D}$. Suppose that $R_{\mathcal{Z}_{2D}} \leq \kappa_2$. Now $OPT_{2D}(s) \geq OPT_{3D}(s)$, so we have

$$R_{\mathcal{Z}_{3D}}(s) = \frac{\mathcal{Z}(s)}{OPT_{3D}(s)} \leq \frac{\mathcal{Z}(s)}{OPT_{2D}(s)} = R_{\mathcal{Z}_{2D}}(s) \leq \kappa_2.$$

Thus $R_{\mathcal{Z}_{3D}}(s) \leq \kappa_2$ for all protein instances s, which implies that $R_{\mathcal{Z}_{3D}} \leq \kappa_2$. Similarly, suppose that $R_{\mathcal{Z}_{2D}} \geq \kappa_1$. Now $OPT_{3D}(s) \geq 20OPT_{2D}(s)$, so we have

$$R_{\mathcal{Z}_{3D}}(s) = \frac{\mathcal{Z}(s)}{OPT_{3D}(s)} \geq \frac{\mathcal{Z}(s)}{20OPT_{2D}(s)} = \frac{1}{20}R_{\mathcal{Z}_{2D}}(s) \geq \frac{\kappa_1}{20}.$$

Thus $R_{\mathcal{Z}_{3D}}(s) \geq \kappa_1/20$ for all protein instances s, which implies that $R_{\mathcal{Z}_{3D}} \geq \kappa_1/20$.

We now prove the asymptotic performance guarantees for $\mathcal{Z}_{3D}$. From Theorem 1 we know that

$$R^N_{\mathcal{Z}_{3D}} = \inf\left\{\frac{\mathcal{Z}(s)}{OPT_{3D}(s)} \mid s \in S^{3D}_N\right\} \geq \inf\left\{\frac{\mathcal{Z}(s)}{20OPT_{2D}(s)} \mid s \in S^{3D}_N\right\}.$$

Now $S^{2D}_N \subseteq S^{3D}_N$ since $OPT_{2D}(s) \geq OPT_{3D}(s)$. Further, we have

$$\inf\{OPT_{2D}(s) \mid s \in S^{3D}_N - S^{2D}_N\} \geq \inf\{OPT_{2D}(s) \mid s \in S^{2D}_N\}$$

from the definition of S^{2D}_N and S^{3D}_N. It follows that

$$R^N_{\mathcal{Z}_{3D}} \geq \inf\left\{\frac{\mathcal{Z}(s)}{20OPT_{2D}(s)} \mid s \in S^{2D}_N\right\} = \frac{1}{20}R^N_{\mathcal{Z}_{2D}},$$

so

$$R^{\infty}_{\mathcal{Z}_{3D}} = \sup_{N \in \mathbf{Z}} R^N_{\mathcal{Z}_{3D}} \geq \sup_{N \in \mathbf{Z}} \frac{1}{20}R^N_{\mathcal{Z}_{2D}} = \frac{1}{20}R^{\infty}_{\mathcal{Z}_{2D}} \geq \kappa_3/20.$$

Now suppose that $R^{\infty}_{\mathcal{Z}_{2D}} \leq \kappa_4$. Then $R^N_{\mathcal{Z}_{2D}} \leq \kappa_4$ for all N. Given $\epsilon > 0$, let s_ϵ be a sequence in S^{2D}_N for which $\mathcal{Z}(s_\epsilon)/OPT_{2D}(s_\epsilon) \leq \kappa_4 + \epsilon$. Such a sequence is guaranteed to exist from the definition of $R^N_{\mathcal{Z}_{2D}}$. Since $S^{2D}_N \subseteq S^{3D}_N$, we have

$$R^N_{\mathcal{Z}_{3D}} \leq \frac{Z(s_\epsilon)}{OPT_{3D}(s_\epsilon)} \leq (\kappa_4 + \epsilon)\frac{OPT_{2D}(s_\epsilon)}{OPT_{3D}(s_\epsilon)} \leq \kappa_4 + \epsilon.$$

Thus we have

$$R^{\infty}_{\mathcal{Z}_{3D}} = \sup_{N \in \mathbf{Z}} R^N_{\mathcal{Z}_{2D}} \leq \lim_{\epsilon \to 0+} \kappa_4 + \epsilon = \kappa_4.$$

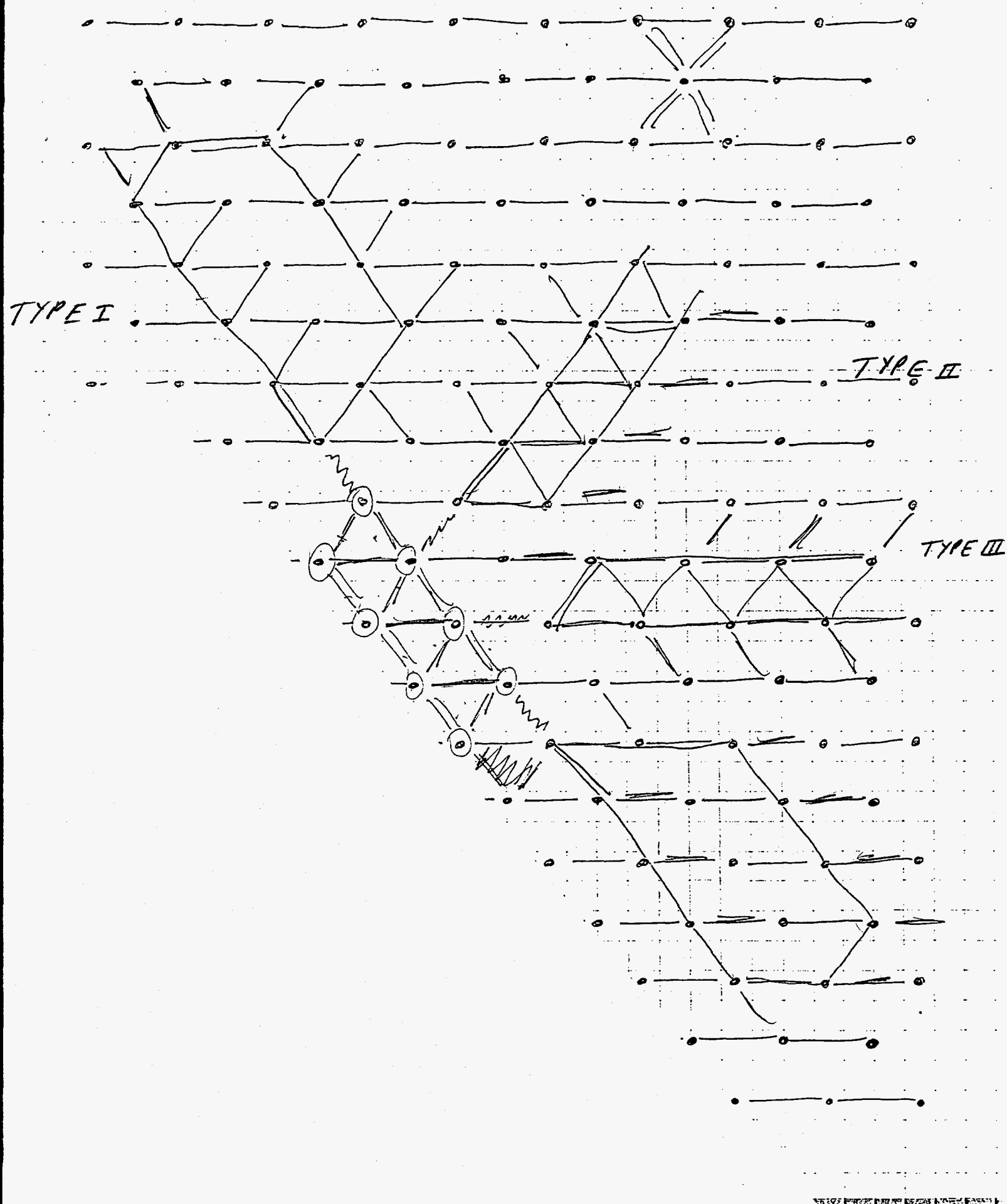
■

Proposition 5 If $\kappa_1 \leq R_{\overline{\mathcal{Z}}_{2D}} \leq \kappa_2$ and $\kappa_3 \leq R^{\infty}_{\overline{\mathcal{Z}}_{2D}} \leq \kappa_4$
then $\kappa_1/10 \leq R_{\mathcal{Z}_{3D}} \leq \kappa_2$ and $\kappa_2/10 \leq R^{\infty}_{\mathcal{Z}_{3D}} \leq \kappa_4$.

Proof. Let $\overline{S}^{2D}_N = \{s \mid \overline{OPT}_{2D}(s) \leq N\}$. Note that $\overline{OPT}_{2D}(s) \geq OPT_{3D}(s)$. It follows that $\overline{S}^{2D}_N \subseteq S^{3D}_N$. Given this observation, the proof of this proposition is analogous to the proof of Proposition 4. ■

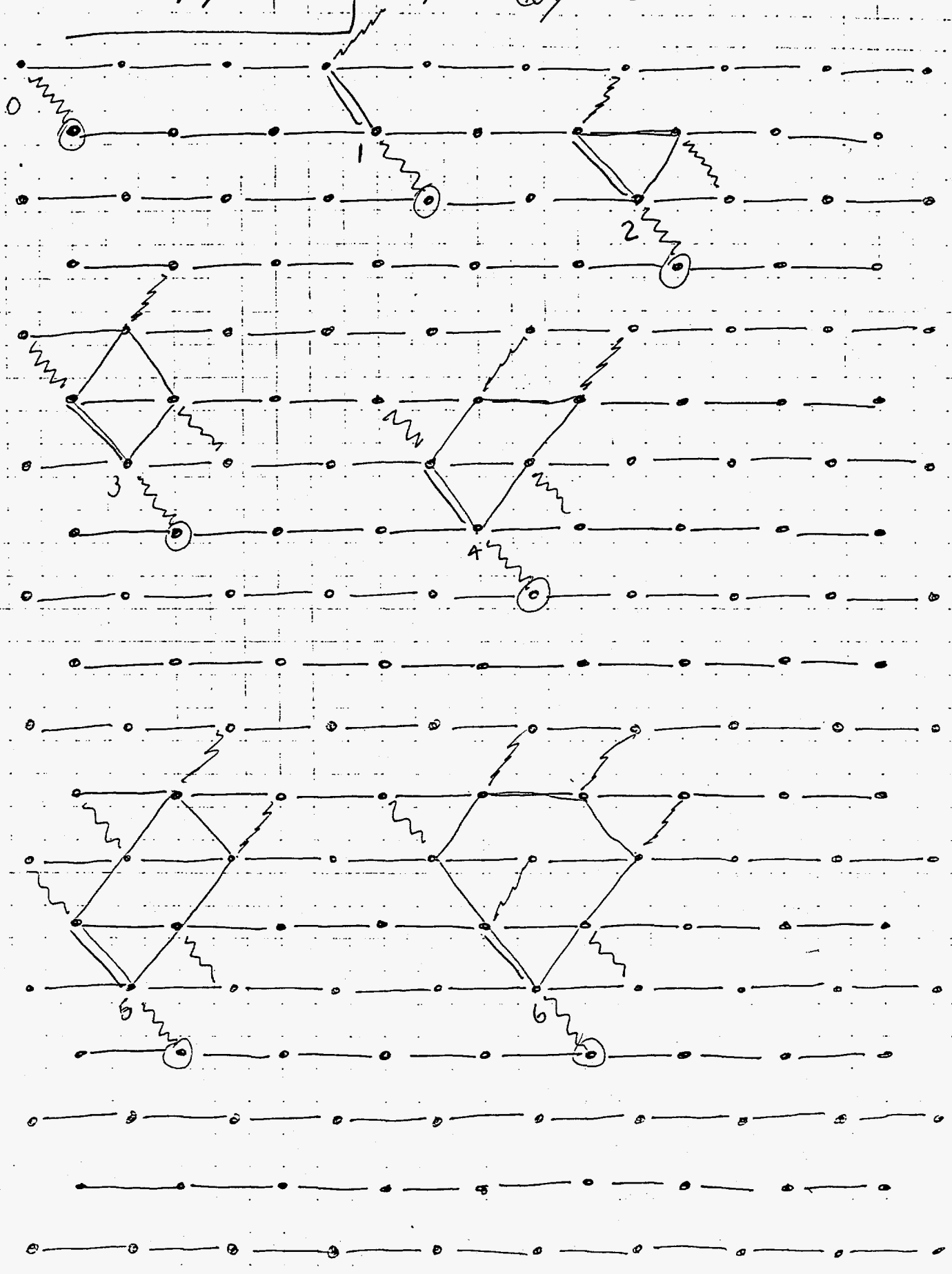
D Illustrations of the hydrophilic loops for Algorithm $\mathcal{C}$

A Single "Layer"

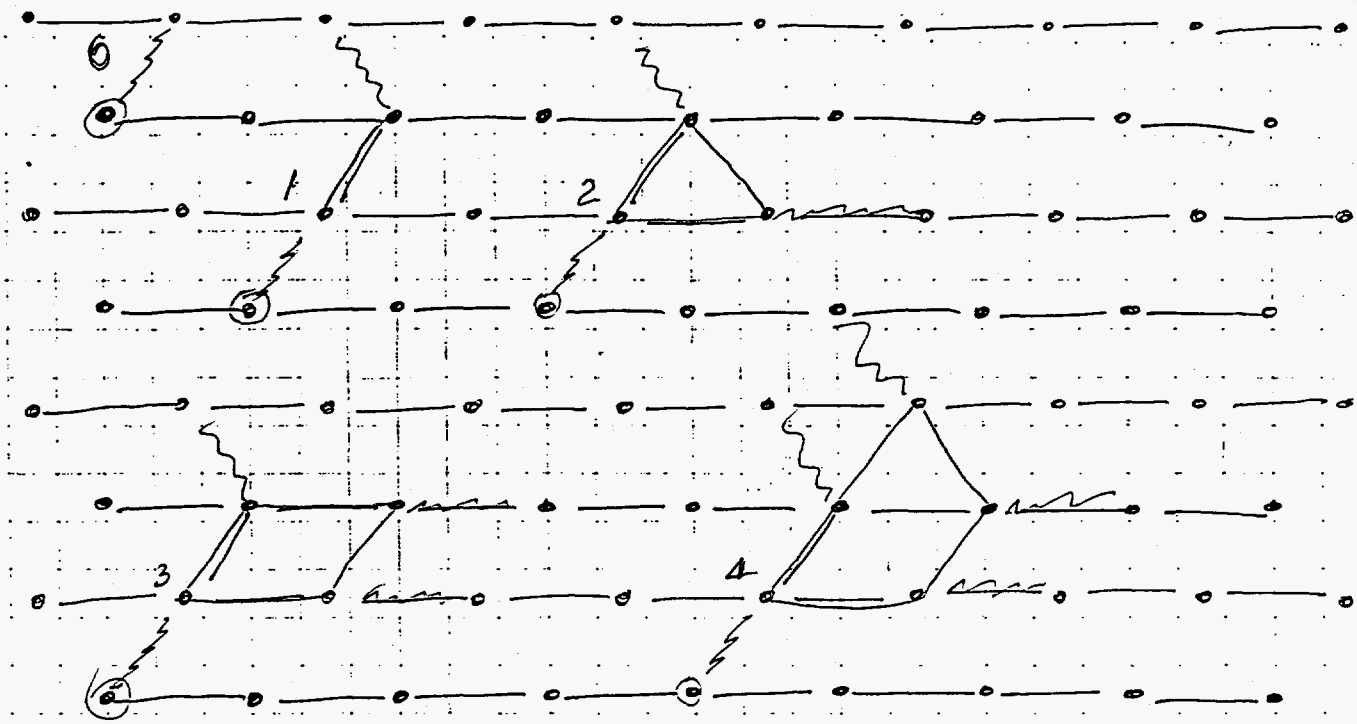


TYPE I

Fat Guy II

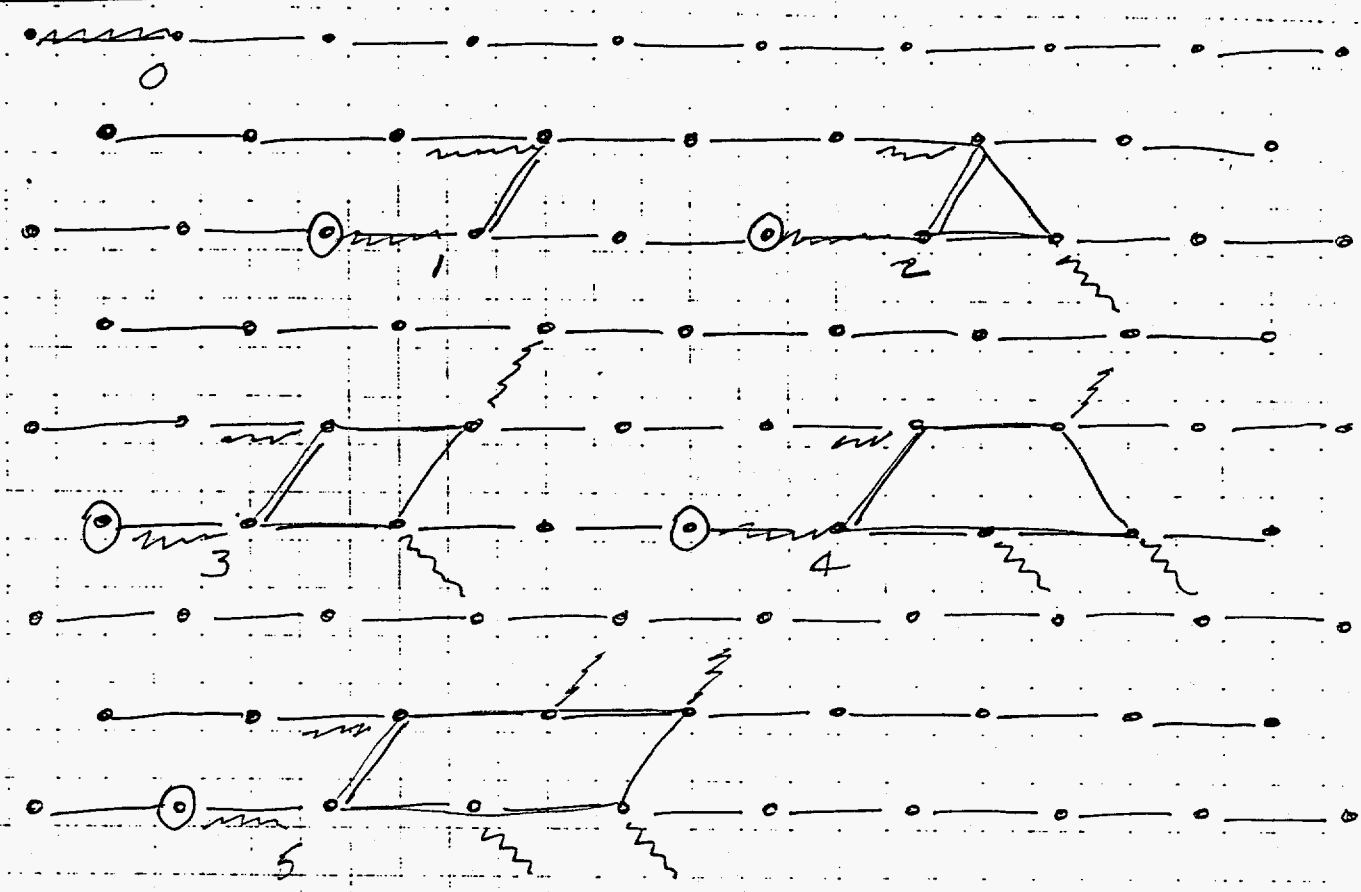


TYPE II Normal Type



TYPE III

Bounded Type



// - edge to next (lower) layer