1

The Computational Complexity of Protein Structure Prediction in Simple Lattice Models

William E. Hart
Sandia National Laboratories

Alantha Newman
Massachusetts Institute of Technology

1.1 Introduction 1-1
1.2 Hydrophobic-Hydrophilic Lattice Models 1-3
1.3 Computational Intractability 1-5
Initial Results • Robust Results • Finite-Alphabet Results
1.4 Performance-Guaranteed Approximation Algorithms 1-8
HP Model • HP Model with Side-Chains • Off-Lattice HP Model • Robust Approximability for HP Models on General Lattices • Accessible Surface Area Lattice Model
1.5 Exact Methods 1-18
Enumeration of Hydrophobic Cores • Constraint Programming • Integer Programming
1.6 Conclusions 1-21

1.1 Introduction

A protein is a complex biological macromolecule composed of a sequence of amino acids. Proteins play key roles in many cellular functions. Fibrous proteins are found in hair, skin, bone, and blood. Membrane proteins are found in cells' membranes, where they mediate the exchange of molecules and information across cellular boundaries. Water-soluble globular proteins serve as enzymes that catalyze most cellular biochemical reactions.

Amino acids are joined end-to-end during protein synthesis by the formation of peptide bonds (see Figure 1.1). The sequence of peptide bonds forms a "main chain" or "backbone" for the protein, off of which project the various side chains. Unlike the structure of other biological macromolecules, proteins have complex, irregular structures. The sequence of residues in a protein is called its *primary structure*. Proteins exhibit a variety of motifs that reflect common structural elements in a local region of the polypeptide chain: α-helices, β-strands, and loops—often termed *secondary structures*. Groups of these secondary structures usually combine to form compact globular structures, which represent the three-dimensional *tertiary structure* of a protein.

The functional properties of a protein depend on its three-dimensional structure. Protein structure prediction (PSP) is therefore a fundamental challenge in molecular biology.

0-8493-8597-0/01/$0.00+$1.50
© 2001 by CRC Press, LLC

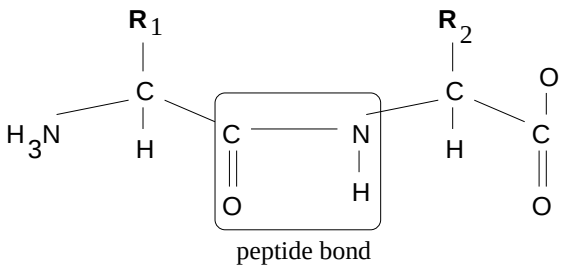

FIGURE 1.1: The peptide bond joining two amino acids when synthesizing a protein.

Despite the fact that the structures of thousands of different proteins have been determined [10], protein structure prediction in general has proven to be quite difficult. The central dogma of protein science is that the primary structure of a protein determines its tertiary structure. Although this is not universally true (e.g. some proteins require *chaperone* proteins to facilitate their folding process), this dogma is tacitly assumed for most of the computational techniques used for predicting and comparing the structure of globular proteins.

Many computational techniques have been developed to predict protein structure, but few of these methods are rigorous techniques for which mathematical guarantees can be described. Most PSP methods employ enumeration or search strategies, which may require the evaluation of exponentially many protein structures. This observation has led many researchers to ask if PSP problems are inherently intractable.

Lattice models have proven to be extremely useful tools for reasoning about the complexity of PSP problems. By sacrificing atomic detail, lattice models can be used to extract essential principles, make predictions, and unify our understanding of many different properties of proteins [18]. One of the important approximations made by lattices is the discretization of the space of conformations. While this discretization precludes a completely accurate model of protein structures, it preserves important features of the problem of computing minimum energy conformations. For example, the related search problem remains difficult and preserves essential features of the conformational space. Consequently, methods that generate low-energy conformations of proteins for lattice models provide insight into the protein folding process.

In this paper, we review results developed in the past decade that rigorously address the computational complexity of protein structure prediction problems in simple lattice models. We consider analyses of (1) intractability, (2) performance-guaranteed approximations and (3) methods that generate exact solutions, and we describe how the lattice models used in these analyses have evolved. Early mathematical analyses of PSP lattice models considered abstract formulations that had limited practical impact, but subsequent work has led to results that (a) apply to more detailed models, (b) consider lattices with greater degrees of freedom, (c) demonstrate the robustness of intractability and approximability, and (d) solve problems with general search frameworks. Our discussion complements the recent review by Chandru et al. [13], who more briefly survey this literature but provide more mathematical detail concerning some of the results in this area.

We begin by describing the the hydrophobic-hydrophilic model (HP model) [17, 29], which is one of the most extensively studied lattice models. Next, we review a variety of results that explore the possible computational intractability of PSP using techniques from computational complexity theory. These results show that the PSP problem is NP-hard in many simple lattice models, and thus widely believed to be intractable. Because

of these hardness results, efficient performance-guaranteed approximation algorithms have been developed for the PSP problem in several lattice models. In particular, many variants of the HP model have been considered, allowing for different degrees of hydrophobicity, explicit side chains and different lattice structures. Finally, we summarize recent efforts to develop exact protein structure prediction methods that provably guarantee that optimal (or near-optimal) structures are found. Although enumerative search methods have been employed for many years, mathematical programming techniques like integer programming and constraint programming offer the possibility of generating optimal protein structures for practical protein sequences.

1.2 Hydrophobic-Hydrophilic Lattice Models

The discretization of the conformational space implicit in lattice models can be leveraged to gain many insights into the protein folding process [18]. For example, the entire conformational space can be enumerated, enabling the study of the folding code. This discretization also provides mathematical structure that can be used to analyze the computational complexity of PSP problems.

A lattice-based PSP model represents conformations of proteins as non-overlapping embeddings of the amino-acid sequence in the lattice. Lattice models can be classified based on the following properties:

1. The *physical structure*, which specifies the level of detail at which the protein sequences are represented. The structure of the protein is treated as a graph whose vertices represent components of the protein. For example, we can represent a protein with a linear-chain structure [18] that uses a chain of beads to represent the amino acids.
2. The *alphabet* of types of amino acids that are modelled. For example, we could use the 20 naturally occurring types of amino acids, or a binary alphabet that categorizes amino acids as hydrophobic (non-polar) or hydrophilic (polar).
3. The *set of protein sequences* that are considered by the model. The set of naturally occurring proteins is clearly a subset of the set of all amino acid sequences, so it is natural to restrict a model to similar subsets.
4. The *energy formula* used, which specifies how pairs of amino acid residues are used to compute the energy of a conformation. For example, this includes contact potentials that only have energy between amino acids that are adjacent on the lattice, and distance-based potentials that use a function of the distance between points on the lattice. Many energy formulas have energy parameters that can be set to different values to capture different aspects of the protein folding process.
5. The *lattice*, in which protein conformations are expressed; this determines the space of possible conformations for a given protein. For example, the cubic and diamond lattices have been used to describe protein conformations (see Figure 1.2).

One of the most studied lattice models is the HP model [17, 29]. This lattice model simplifies a protein's primary structure to a linear chain of beads. Each bead represents an amino acid, which can be one of two types: H (hydrophobic, i.e. nonpolar) or P (hydrophilic, i.e. polar). This model abstracts the hydrophobic interaction, one of the dominant forces in protein folding. Although some amino acids are not hydrophilic or hydrophobic in all contexts, the model reduces a protein instance to a string of H's and P's that represents the pattern of hydrophobicity in the protein's amino acid sequence. Despite its simplicity, the

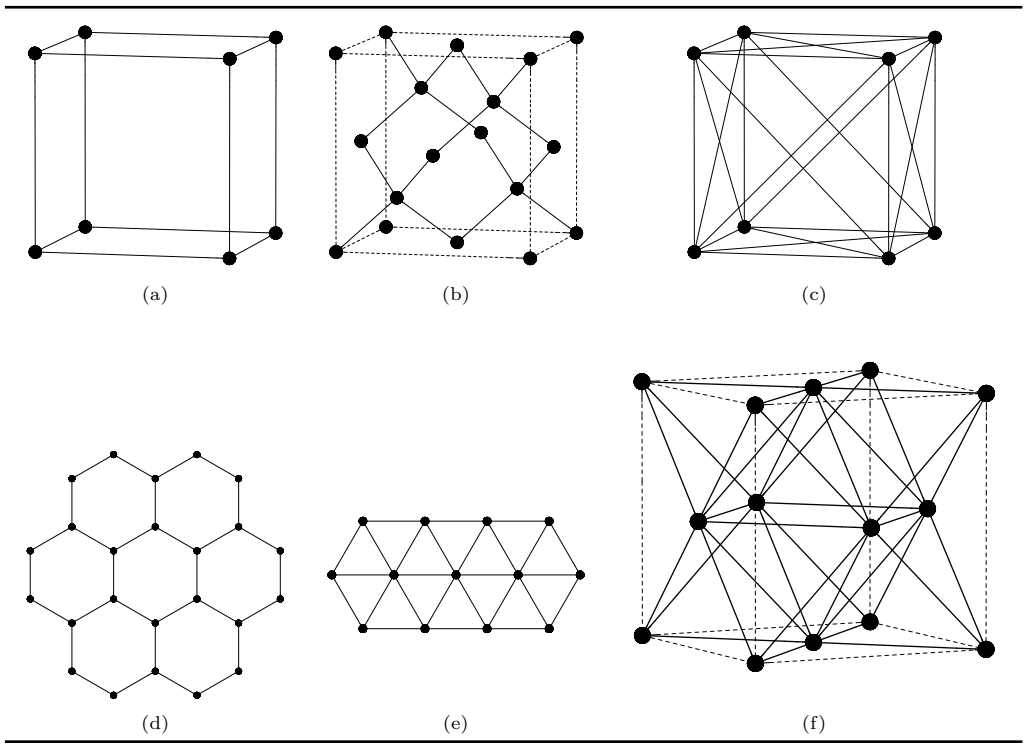

FIGURE 1.2: Examples of crystal lattices: (a) cubic, (b) diamond, (c) cubic with planar diagonals, (d) hexagonal, (e) triangular and (f) face-centered-cubic.

model is powerful enough to capture a variety of properties of actual proteins and has been used to discover new properties. For example, proteins in this model collapse to compact states with hydrophobic cores and significant amounts of secondary and tertiary structure.

For simplicity, we denote H by "1" and P by "0", so the alphabet used in an HP model is $A = \{0, 1\}$. The set of protein instances typically considered for this model is the set of binary sequences $\sigma = \{0, 1\}^+$. Each sequence $s \in \sigma$ corresponds to a (hypothesized) hydrophobic-hydrophilic pattern of a protein sequence. The HP model uses contact energies between pairs of amino acids: two amino acids can contribute to the protein's energies if they lie on adjacent points in the lattice. Thus the energy formula used in the HP model is an energy matrix, $\mathcal{E} = (e(a, b))_{a,b \in A}$, where $e(a, b) = -1$ if $a = b = 1$, and $e(a, b) = 0$ otherwise. The HP model studied by Dill and his colleagues models protein conformations as linear chains of beads folded in the 2D square or 3D cubic lattices.

Much of our review of the computational complexity of PSP focuses on the HP model, because it has been so widely studied. Additionally, a variety of extensions of the HP model have been considered in an effort to make these PSP results more practically relevant. For example, Agarwala et al. [1] consider an extension of the HP model that allows for various degrees of hydrophobicity.

More general structures have also been considered than the standard linear-chain model. One example is a simple side-chain structure that uses a chain of beads to represent the backbone; amino acids are represented by beads that connect to a linear backbone with a

single edge [11, 25, 28]. Figure 1.3 contrasts the structure of linear and side-chain conformations in the HP model.

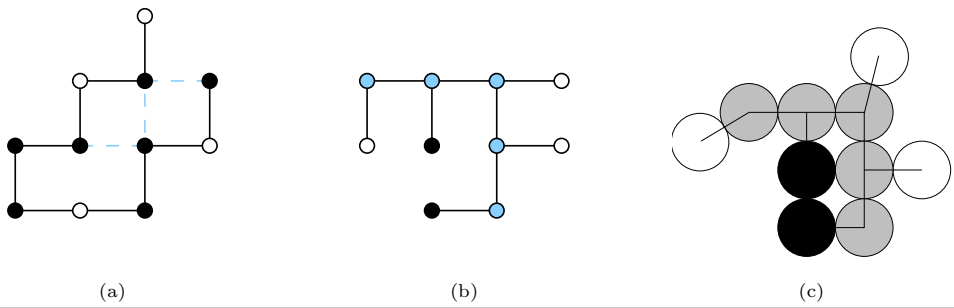

FIGURE 1.3: Illustrations of conformations for: (a) the standard HP model on the square lattice, (b) the HP model with side chains on the square lattice, and (c) the HP tangent spheres model with side chains. Black denotes a hydrophobic amino acid, white denotes a hydrophilic amino acid, and gray denotes a backbone element.

Although most work on the HP model has focused on the 2D square and 3D cubic lattices, the computational complexity of PSP for the HP model has been studied for a variety of different lattices, including the triangular lattice (see Figure 1.2(e)) [1], the face centered cubic (FCC) lattice (see Figure 1.2(f)) [1, 25], the cubic lattice with diagonal edges on each face (see Figure 1.2(c)) [28], and general crystallographic lattices [27]. Off-lattice variants of the HP model have also been explored by treating a protein structure as a set of connected spheres, with a contact interaction potential that is identical to the standard HP model (see Figure 1.3(c)) [25]. The term off-lattice is used because the protein is not actually folded on a lattice. Conformations on a given lattice can clearly be translated into conformations in this off-lattice model, and near-optimal conformations on triangular and FCC lattices are closely related to near-optimal off-lattice conformations.

1.3 Computational Intractability

Exhaustive search of a protein's conformational space is clearly not a feasible algorithmic strategy. The number of possible conformations is exponential in the length of the protein sequence, and powerful computational hardware would not be capable of searching this space for even moderately large proteins. This observation led Levinthal to raise a question about the paradoxical discrepancy between the enormous number of possible conformations and the fact that most proteins fold within seconds to minutes [36]. While these observations appear contradictory, they can be reconciled by noting that they may simply point to the lack of knowledge that could be used to design an efficient search algorithm (see Ngo et al. [36] for further discussion of this issue). Computational analyses of PSP address this lack of knowledge by providing insight into the inherent algorithmic difficulty of folding proteins.

The native conformation of a protein is the conformation that determines its biological function. Following the thermodynamic hypothesis [19], computational models of protein folding are typically formulated to find the global minimum of a potential energy function. In lattice models, an energy value is associated with every conformation taking into account particular neighborhood relationships of the amino acids on the lattice. Consequently, given a lattice model L and sequence s, the PSP problem is to find a conformation of s in L with minimal energy.

Computational intractability refers to our inability to construct efficient (i.e., polynomial-time) algorithms that can solve a given problem. Here, "inability" refers to both the present state-of-the-art of algorithmic research as well as possible mathematical statements that no such algorithms exist. Customary statements about the intractability of a problem are made by showing that the problem is NP-hard. It is widely believed that a polynomial-time algorithm does not exist for any NP-hard problem, since the class of NP-hard optimization problems includes a wide variety of notoriously difficult combinatorial optimization problems. The best known algorithm for any NP-hard problem requires an exponential number of computational steps, which makes these problems "practically intractable."

1.3.1 Initial Results

PSP has been shown to be NP-hard for various lattice models. Initial intractability analyses of PSP considered models that captured PSP problems in rather limited and unrealistic ways. We survey these analyses and then critique these PSP results in the following two sections.

Fraenkel [20] present a NP-hardness results for a physical model in which each amino acid is represented as a bead connected to a backbone. The protein must be embedded in a cubic lattice subject to pairwise constraints on the beads, i.e. specified pairs of beads, including pairs of beads on the backbone, are required to be at a fixed distance in the embedding. These specified pairs comprise a *contact graph*. The alphabet consists of three types that represent the charges associated with the amino acids: -1, 0, 1. The model uses a distance-dependent energy formula that computes the product of the charges divided by distance. The energy is the sum over all edges in the contact graph.

Ngo and Marks [35] present a NP-hardness result for a molecular structure prediction problem that encompasses protein structures. This model considers a chain molecule of atoms that is to be embedded in a diamond lattice. The energy formula is based upon a typical form of the empirical potential-energy function for organic molecules, which is a distance-dependent function.

Paterson and Przytycka [37] present a NP-hardness result for a physical model in which each amino acid is represented as a bead along a chain that is to be embedded in a cubic lattice. A contact energy formula is used, so a pair of amino acids contributes to the conformational energy only if they are adjacent on the lattice. This energy formula has contact energies of one for contacts between identical residues and zero otherwise. The amino acid types in this model are not limited a priori, so instances of this model can represent instances of many specific contact-based PSP problems. However, we note below that this generality is actually a weakness of the model.

Finally, Unger and Moult [43] present a NP-hardness result for a physical model in which each amino acid is represented as a bead along a chain that is to be embedded in a cubic lattice with planar diagonals. The energy formula is a simple form of the empirical potential energy-function for organic molecules, which is a distance-dependent calculation. This NP-hardness result can be generalized to the Bravais lattices (which includes the cubic lattice), as well as the diamond and fluorite lattices [24].

1.3.2 Robust Results

It is difficult to provide strong recommendations for particular PSP formulations because accurate potential energy functions are not known. While various analytic formulations use potentials that capture known features of "the" potential function, the most appropriate analytic formulation of the potential energy for PSP remains an area of active research [15, 44]. Consequently, robust algorithmic results are particularly important for lattice-based PSP models.

Computational robustness refers to the independence of algorithmic results from particular settings. In the context of NP-completeness, robustness refers to the fact that a class of closely related problems can be described, all of which are NP-complete. The members of the class of problems are typically distinguished by some parameter(s) that form a set of reasonable alternate formulations of the same basic problem. Intractability results for PSP can be robust in two different ways [26]. First, an intractability result can be robust to changes in the lattice. The analysis of the PSP problem formulation posed by Unger and Moult [43], which uses a simplified empirical energy potential, can be generalized to show that this PSP problem is NP-hard for any finitely representable lattice [26].

Second, an intractability result can be robust to changes in the energy. Consider a PSP formulation with an objective of the form

$$\sum_{i=2}^{n} \sum_{j=1}^{i-1} C_{s_i,s_j} g\left(|f_i - f_j|\right), \tag{1.1}$$

where $g : \mathbf{Q} \to \mathbf{R}$ is an energy potential that monotonically increases to zero (in an inversely quadratic fashion) as the distance between amino acids increases. This model can be viewed as a special case of the model examined by Unger and Moult [43], and the class of functions g includes widely used pairwise potential functions like the Lennard-Jones potential. Additionally, the use of the distance $|f_i - f_j|$ makes this energy formulation translationally invariant, which is consistent with practical emperical energy models. For any function g and for an appropriate discretization of the L_2 norm, this PSP problem is NP-hard [26]. Additionally, this result can be generalized to show that this PSP problem is also NP-hard if the protein is modeled with explicit side-chains instead of as a simple linear chain.

1.3.3 Finite-Alphabet Results

A significant weakness of almost all of the models used in these intractability results is that the alphabet of amino acid types used to construct protein sequences is unbounded in size.[1] Let an amino acid *type* be defined by the pattern of interactions it exhibits with all other amino acids. These PSP problems allow for problem instances for which the number of amino acid types are not bounded. For example, a PSP formulation that uses Equation (1.1) allows for $O(n^2)$ amino acid types because the interaction between amino acids i and j is defined in part by the matrix coefficient C_{s_i,s_j}, which can assume any value.

Consequently, the previous models do not accurately model physically relevant PSP problems, for which there are 20 naturally occurring amino acid types. To address this concern, several authors have developed complexity analyses for models with a finite set of amino

[1] Fraenkel's model [20] uses a finite number of amino acid types, but it allows the protein chain to be embedded in a lattice without forcing subsequent amino acids to lie in close proximity on the lattice, thereby leading to biologically implausible conformations for certain amino acid sequences.

acids. For example, a PSP problem for which protein sequences are defined from a set of 12 amino acid types and the conformational energy is computed using a contact potential was proved to be NP-hard [2]. Nayak, Sinclair and Zwick [32] consider a string folding problem with a very large alphabet of amino acids, using a technique that "converts" a hardness proof for a model with an unbounded number of amino acids to a hardness proof in a model with a bounded number of amino acids. Crescenzi et al. [16] and Berger and Leighton [9] prove that PSP in the simple HP-model is NP-hard for the 2D square and 3D cubic lattices, respectively.

1.4 Performance-Guaranteed Approximation Algorithms

Performance guaranteed approximation algorithms complement intractability analyses by demonstrating that near-optimal solutions can be efficiently computed. An approximation algorithm has a multiplicative asymptotic approximation ratio of α if the solutions generated by the algorithm are within a factor of α of the optimum. Performance guaranteed approximation methods have been developed for a variety of HP lattice models, as well as some natural generalizations of the HP model.

1.4.1 HP Model

Performance guaranteed approximation algorithms have been developed for the HP model on the 2D square lattice, 3D cubic lattice, triangular lattice and the face-centered-cubic (FCC) lattice [1, 23, 31, 33, 34]. These approximation algorithms take an HP sequence $s \in \{0,1\}^+$, and form a conformation on the lattice. Recall that the energy of a conformation is the number of hydrophobic-hydrophobic contacts: hydrophobics (1's) that are adjacent on the lattice but not adjacent on the string.

Square Lattice

The PSP problem in the HP model takes as input an HP sequence S, which can be viewed as a binary string (H=1, P=0). The objective is to find a folding of the string s that forms a self-avoiding walk on a specified lattice and maximizes the number of contacts. Figure 1.4 illustrates an optimal conformation for a binary string on the 2D square lattice (i.e. with the maximum number of contacts). Let $\mathcal{E}[s]$ denote the number of 1's in even positions in the sequence s (even-1's) and let $\mathcal{O}[s]$ denote the number of 1's in odd positions in s (odd-1's). Additionally, let

$$X[s] = \min\{\mathcal{E}[s], \mathcal{O}[s]\}. \tag{1.2}$$

Due to the fact that the square lattice is bipartite, each even-1 in s can have contacts only with odd-1's in s and vice-versa. In any conformation of s on the 2D square lattice, each 1 in the string s that is not in the first or last position on the string can have at most two contacts. Thus, an upper bound on the maximum number of contacts in any conformation of s on the 2D square lattice is:

$$2 \cdot X[s] + 2. \tag{1.3}$$

The first approximation algorithm developed for the PSP problem on the square lattice has an approximation ratio of 1/4 [23]. For a given sequence s, this algorithm first finds a point p in s such that at least half the odd-1's are in one substring on one side of p (the *odd* substring) and at least half the even-1's are on the other side of p (the *even*

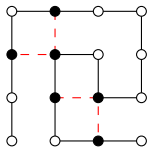

FIGURE 1.4: An optimal conformation for the string 0010100001011010 on the 2D square lattice. This conformation has four contacts.

substring). Then, the odd substring is embedded in the square lattice such that all odd-1's in the odd substring have the same y-coordinate and the even substring is embedded in a complementary fashion (see Figure 1.5). This conformation yields at least $X[s]/2$ contacts, which is 1/4 of optimal. Mauri, Piccolboni and Pavesi [31] describe an algorithm that also has an approximation ratio of 1/4, which they argue works better in practice.

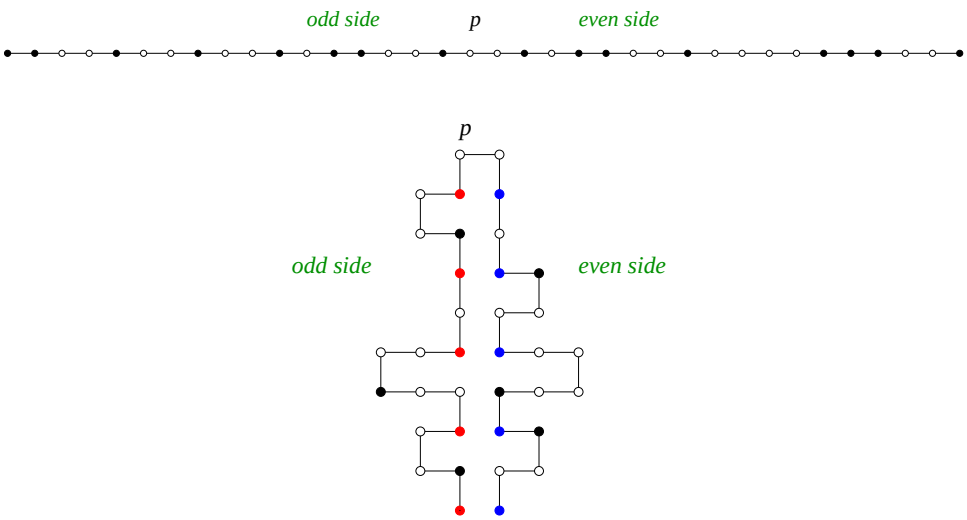

FIGURE 1.5: Illustration of a conformation generated by a simple 1/4-approximation algorithm for the HP model on the square lattice.

The approximation ratio for this problem can be improved to 1/3 [33]. For simplicity, we consider even-length sequences s for which $\mathcal{O}[s] = \mathcal{E}[s]$. This approximation algorithm creates conformations as folded loops, where the end-points are adjacent on the lattice. For example, the conformation in Figure 1.5 can be viewed as a folded loop. The first step in the algorithm is to find a point p such that as we move clockwise in the loop starting at point p, we encounter at least as many odd-1's as even-1's and as we go counter-clockwise, we encounter at least as many even-1's as odd-1's.

Let $B_{\mathcal{O}}$ be the distance between the first pair of consecutive odd-1's encountered as we go in the clockwise direction starting at point p and let $B_{\mathcal{E}}$ be the distance between the first pair of consecutive even-1's encountered as we go in the counter-clockwise direction. We sketch the algorithm in Figure 1.6. In cases (a) and (b) of Step 2, we form three contacts

and use at most four even- and odd-1's and "waste" at most four even- and odd-1's, i.e. we waste even-1's that occur on the odd side and vice-versa. In cases (c) and (d), we form two contacts and use at most three even- and odd-1's and waste at most three even- and odd-1's. Since there are at most $2\mathcal{O}[s] + 2 = \mathcal{O}[s] + \mathcal{E}[s] + 2$ contacts, this gives a 1/3 approximation ratio.

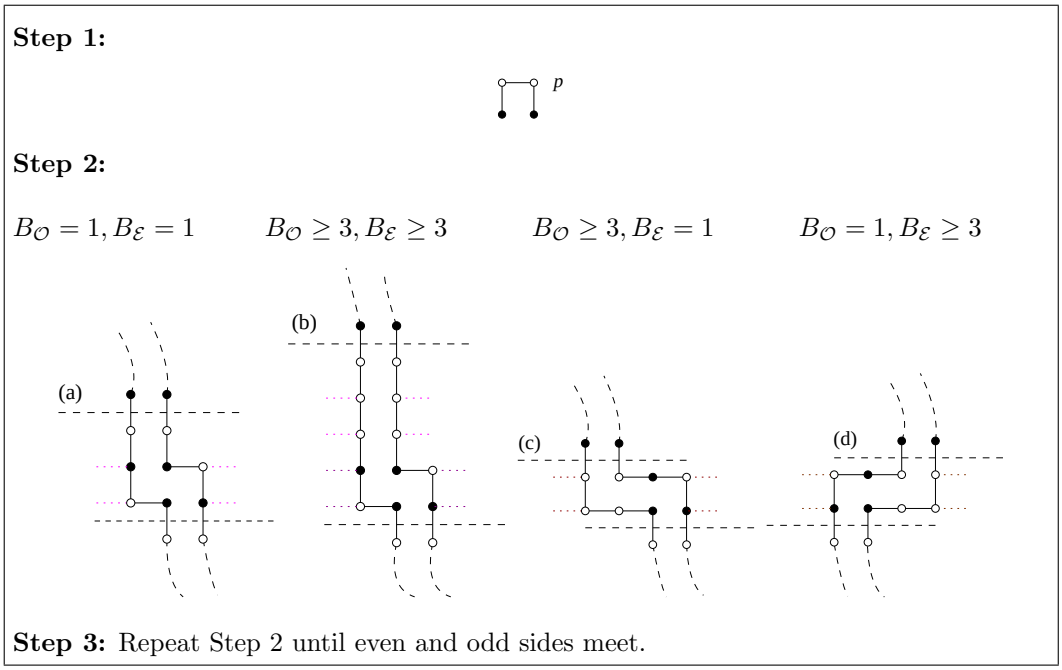

FIGURE 1.6: The steps used in the 1/3-approximation algorithm for the folding problem in the HP model on the square lattice.

Cubic Lattice

In any folding of a sequence s on the 3D cubic lattice, each 1 in the string s that is not in the first or last position can have at most four contacts. Thus, an upper bound on the maximum number of contacts in any conformation of s on the 3D cubic lattice is:

$$4 \cdot X[s] + 2. \tag{1.4}$$

The 1/4-approximation algorithm described above can be generalized to an approximation algorithm for the problem on the 3D cubic lattice [23]. Suppose the odd side of s has at least k odd-1's and the even side has at least k even-1's, i.e. $k \geq X[s]/2$. Then we can divide the odd side into segments with $\sqrt{k}$ odd-1's and divide the even side into segments with $\sqrt{k}$ even-1's. This approximation algorithm repeats the 2D folding algorithm $\sqrt{k}$ times in adjacent planes, i.e. the first pair of segments is folded in the plane $z = 0$, then next in the plane $z = 1$, etc. In the resulting conformation, each of $X[s]/2 - c\sqrt{X[s]}$ odd-1's has at least 3 contacts for some constant c. Thus, this algorithm has an approximation ratio of $3/8 - \Omega(1/\sqrt{X[s]})$.

Another approximation algorithm, based on different geometric ideas, improves on this absolute approximation guarantee [34]. In this algorithm, the string s is divided into two substrings so that one substring contains at least half the odd-1's and the other substring at least half the even-1's. Each substring is folded along two different diagonals, as shown in Figure 1.7. All but a constant number of odd-1's from the odd substring get three contacts. These geometric ideas can be used to obtain a slightly improved approximation ratio of .37501, which shows that 3/8 is not the best approximation guarantee that can be obtained for this problem, despite the fact that it was the best guarantee known for the past decade.

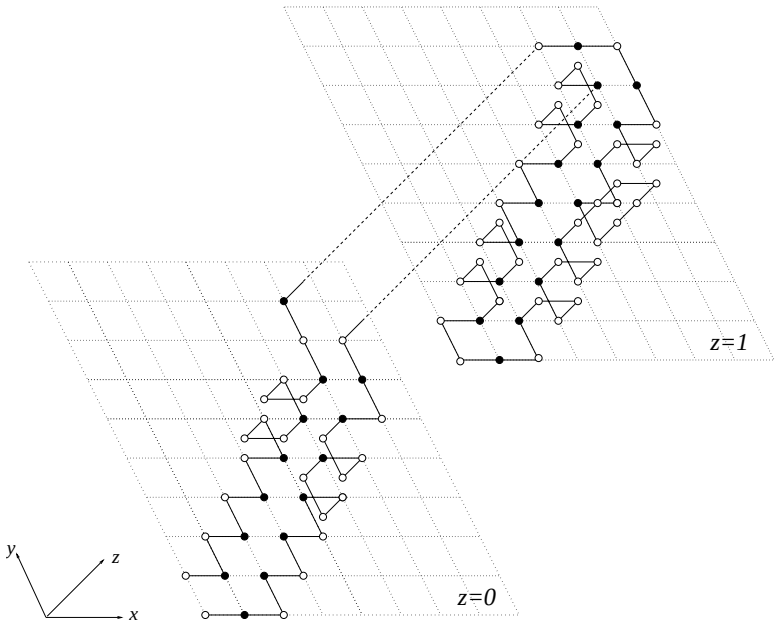

FIGURE 1.7: An illustration of a conformation generated by folding substrings along diagonals of the cubic lattice.

Triangular and FCC Lattices

One undesirable feature of the square lattice is that a contact must be formed between hydrophobics with different parities. There is no such parity restriction in real protein folds. This issue is discussed by Agarwala et al. [1], who suggest that the triangular lattice is more suitable to model protein folding. They give simple 1/2-approximation algorithms and a 6/11-approximation algorithm that uses an improved upper bound. Agarwala et al. generalize these results to a 3D triangular lattice that is equivalent to the FCC lattice, for which they describe an algorithm with an approximation ratio of 3/5.

1.4.2 HP Model with Side-Chains

Performance guaranteed approximation algorithms have also been developed for an HP model that explicitly represents side chains [26, 28]. This lattice model represents the

conformation of a protein using a subclass of branched polymers called "branched combs." A homopolymer version of this model was introduced by Bromberg and Dill [11], who argued that linear lattice models fail to capture properties of protein folding, like side chain packing, that affect the stability of the native protein structure. The HP side chain model 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 side chain beads is labelled hydrophobic or hydrophilic.

Figure 1.3(b) illustrates a conformation of the HP side chain model on the square lattice. Note that there are no interactions between backbone elements and side-chain elements, so the energy of such a conformation is simply the number of contacts between hydrophobic side chains on the lattice. Further, note that adjacent side chains can contribute energy in this model, which is a fundamental difference induced by the branched combs structure.

Figure 1.8 illustrates the repeated conformational structure produced by an approximation algorithm for the problem on the square lattice [25]. The folding point for this algorithm is selected in the same manner as for the linear chain model, and thus this structure can be constructed in linear time. This algorithm guarantees that for a string s, $\lfloor X[s]/4 \rfloor$ hydrophobic-hydrophobic contacts are formed between the two halves of the conformation. Since each hydrophobic side chain can have at most three contacts, this algorithm has a 1/12 approximation ratio.

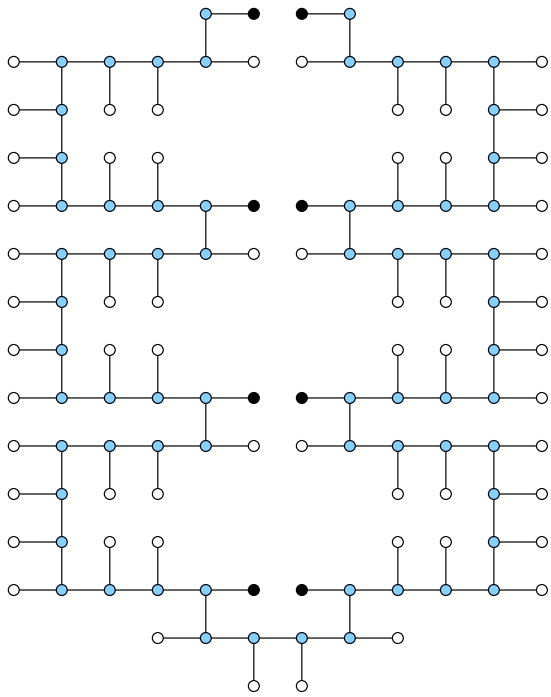

FIGURE 1.8: An illustration of the conformations generated by an approximation algorithm for the HP side chain model on the square lattice.

A similar algorithm for the 3D cubic lattice can also be developed [25]. This approximation algorithm also divides the protein at a folding point, but it then attempts to create a 3D fold with four columns of hydrophobics in the core. Figure 1.9 illustrates the structure of one of these columns, as well as how the protein sequence forms a hydrophobic core. The hydrophobic core is formed by threading each half of the protein sequence through the four columns in an anti-parallel fashion (e.g. up - down - up - down). In this conformation, it contains at least $4\lceil X[s]/2\rceil - 20$ contacts for a sufficiently large sequence s. Since each hydrophobic side chain can have at most five contacts, this algorithm as a 4/10 approximation ratio.

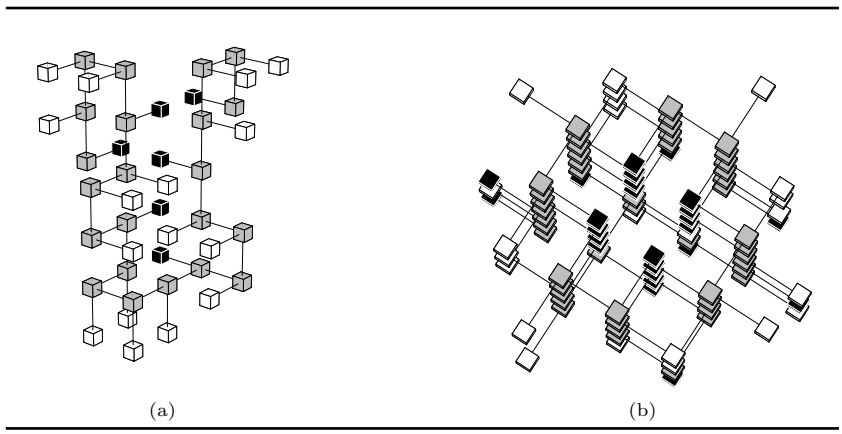

FIGURE 1.9: Illustration of the conformations generated by an approximation algorithm for the HP side chain model on the cubic lattic: (a) the 3D structure of a single column, and (b) a perspective of the core generated by interlacing the columns.

These approximation results have been generalized to lattices that do not have the parity restriction imposed by the cubic lattice: the FCC lattice and the cubic lattice with facial diagonals (which Heun calls the *extended cubic lattice* (ECL)) [25, 28]. Both of these lattices allow any hydrophobic amino acids to be in contact with any other hydrophobic amino acid. Thus if there are $N(s)$ hydrophobic amino acids in a sequence s then we can obtain upper bounds of $9N(s)/2$ contacts for the FCC lattice and $7N(s)$ contacts for the ECL.

These approximation algorithms are very similar in that they both place all hydrophobic side chains in a set of columns, with an algorithm that forms the conformation in a linear fashion (layer by layer or column by column). The hydrophobic columns form a distinct hydrophobic core, with an irregular outer layer of hydrophilic side chains. For example, Figure 1.10 illustrates a conformation generated by an approximation algorithm for the FCC lattice [25], which generates eight columns of hydrophobics. These tight hydrophobic cores guarantee that these approximation algorithms have an approximation ratio of 31/36 on the FCC lattice and 59/70 on the ECL.

Heun [28] also considers approximation algorithms that are tailored to the characteristics of sequences commonly found in the SWISS-PROT protein database. Specifically, Heun considers HP sequences that can be decomposed into blocks of 6 hydrophobics of the form $\sigma = P^{l_1}H \ldots P^{l_6}H$ where

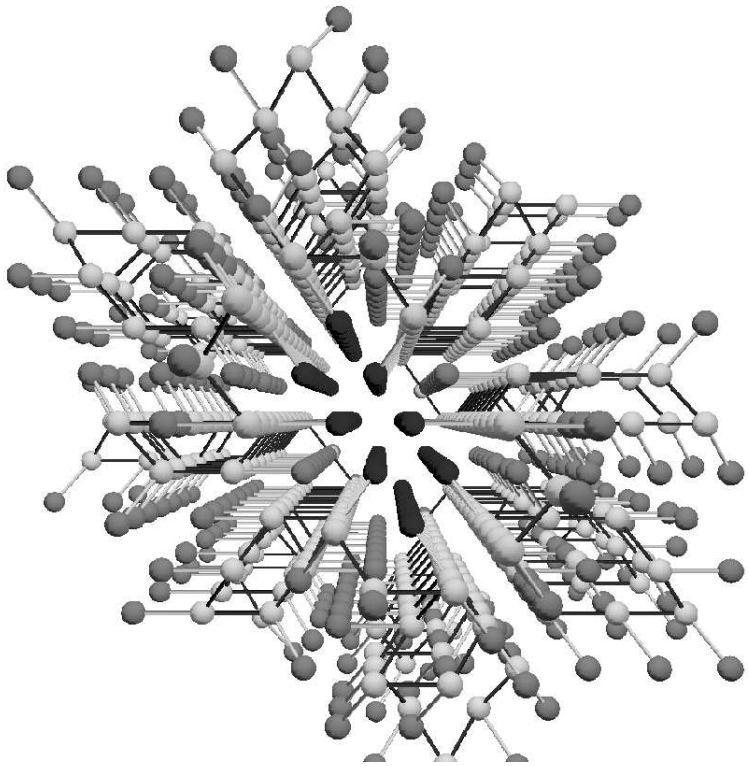

FIGURE 1.10: Illustration of the conformation generated by an approximation algorithm for the HP side chain model on the FCC lattice. The structures generated by Heun's approximation algorithm on the extended cubic lattice have a similar structure, with 10 hydrophobic columns in the core.

- either there exists $i \in \{2, 3, \ldots, 6\}$ such that $l_i = 0$, or
- there exists $i, j \in \{1, 2, \ldots, 6\}$, $i \neq j$, such that $l_i + l_j \leq 3$.

Heun notes that over 96% of the sequences in SWISS-PROT can be decomposed into blocks of 6 hydrophobics with this character, and he describes an approximation algorithm for the ECL with an approximation ratio of 37/42.

1.4.3 Off-Lattice HP Model

The HP tangent spheres models are simple PSP models that do not use a lattice but are analogous to the standard HP model [25]. Because the conformations in these models are not defined within a lattice, these models are termed off-lattice models. In these models, the graph that represents the protein is transformed to a set of tangent spheres of equal radius (or circles in two dimensions). 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 1.3(c)). The linear chain model represents the protein as a sequence of spheres on a string, consecutive spheres being tangent, which are labelled hydrophobic or hydrophilic.

The side chain model represents the backbone as in the linear chain model, but now every sphere in the backbone is tangent to a side chain sphere that models the physical presence of that amino acids side chain. The side chain spheres are labelled hydrophobic or hydrophilic. A hydrophobic-hydrophobic contact in such a model is obtained when two hydrophobic-spheres are tangent.

The tangent spheres side chain model generalizes the HP model in the sense that for any lattice a conformation on that lattice represents a possible off-lattice conformation. Thus HP tangent spheres models can be analyzed rigorously by transferring algorithmic analyses from various lattice HP-models to the off-lattice setting. In 2D, the maximum number of spheres that can be tangent to a single sphere is 6. Thus a hydrophobic sphere in a linear chain can be tangent to at most 4 other hydrophobic spheres. The arrow-folding algorithm described by Agarwala et al. [1] can be used to construct a conformation (with the linear sphere chain) that has at least $N(s) - 3$ hydrophobic-hydrophobic contacts. Consequently, this algorithm has a $1/4$ approximation ratio for the HP tangent spheres model.

To analyze the performance of the HP tangent spheres model in three dimensions, recall 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 3D kissing number. From this we can conclude that a hydrophobic sphere in a linear chain can be tangent to only 10 other hydrophobic spheres, and a hydrophobic side chain sphere in a side chain model can be tangent to only 11 other hydrophobic side chain spheres. Thus each hydrophobic sphere in a linear chain can contribute at most 5 contacts and each hydrophobic side chain can contribute at most $11/2$ contacts.

The star-folding algorithm described by Agarwala et al. [1] can be used to construct a FCC conformation (with the linear sphere chain) that has $8N(s)/3$ hydrophobic-hydrophobic contacts (ignoring boundary conditions). Consequently, this algorithm has a $8/15$ approximation ratio for the HP tangent spheres model. Similarly, the approximation for the FCC side chain model [25] can be used to construct a conformation that has at least 31 $N(s)/8 - 42$ contacts (for sufficiently long sequences). Consequently, this algorithm has a $31/44$ approximation ratio for the HP tangent spheres model with side chains.

1.4.4 Robust Approximability for HP Models on General Lattices

The results that we have surveyed in this section demonstrate that near-optimal protein structures can be quickly constructed for a variety of HP lattice models as well as simple off-lattice protein models. This naturally begs the question of whether approximability is a general property of HP lattice models. Results that transcend particular lattice frameworks are of significant interest because they can say something about the general biological problem with a higher degree of confidence. In fact, it is reasonable to expect that there will exist algorithmic invariants across lattices that fundamentally relate to the protein folding problem, because lattice models provide alternative discretizations of the same physical phenomenon.

Two "master" approximation algorithms have been developed for bipartite and non-bipartite lattices that demonstrate how approximation algorithms can be applied to a wide range of lattices [27]. These master approximation algorithms provide a generic template for an approximation algorithm using only a sublattice called a latticoid, a structured sublattice that in which a skeleton of hydrophobic contacts can be constructed. Further, the analysis of these algorithms includes a complexity theory for approximability in lattices that can be used to transform PSP algorithms in one lattice into PSP algorithms in another lattice such that we can provide a performance guarantee on the new lattice.

Figure 1.11 represents two possible latticoids of the square lattice. The bipartite master

approximation algorithm selects a folding point in the same fashion used for the 2D HP model [23], and a hydrophobic core is similarly made by pairing odd and even hydrophobics along two faces of the conformation. The central row in these latticoids indicates the points at which hydrophobic contacts can be made by the master approximation algorithm.

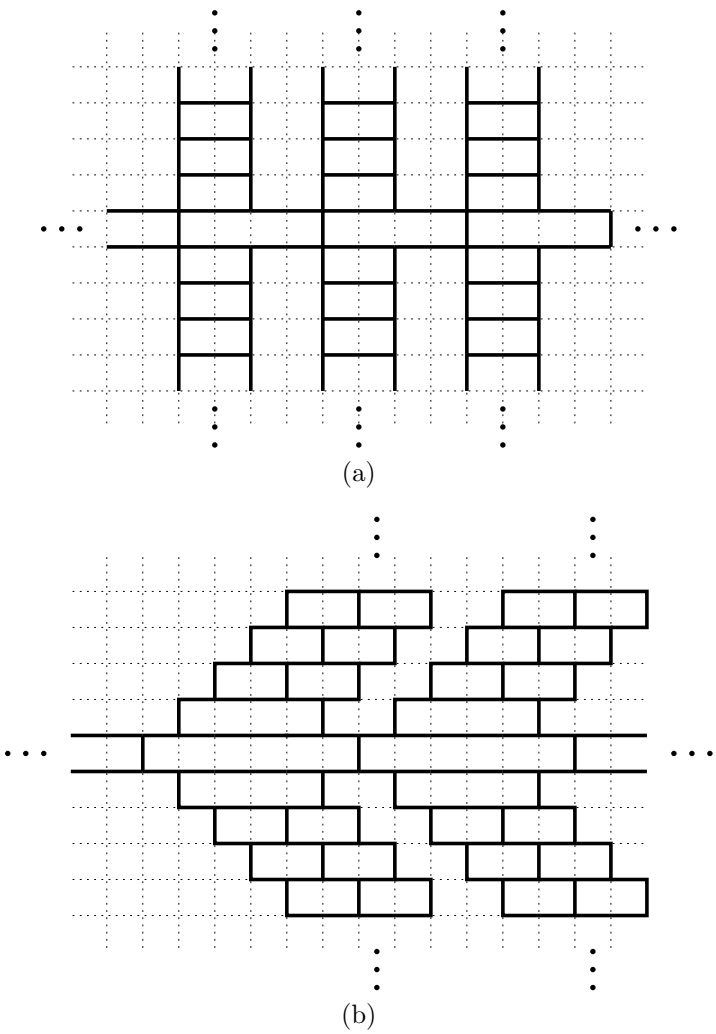

FIGURE 1.11: Illustrations of two latticoids of the square lattice. Dark lines indicate edges that are used in some protein conformation and dashed lines indicate remaining edges in the square lattice. The contact edges are the bolded edges in the central horizontal row.

The latticoids in Figure 1.11 can be embedded into a wide range of crystal lattices to provide a performance guaranteed approximation algorithm for the HP model. To illustrate this, consider the diamond lattice, whose unit cell is shown in Figure 1.2(b). Figure 1.12 illustrates how the latticoid in Figure 1.11(a) can be embedded into this lattice to ensure that at least $\lfloor X[s]/4 \rfloor$ hydrophobic-hydrophobic contacts are formed.

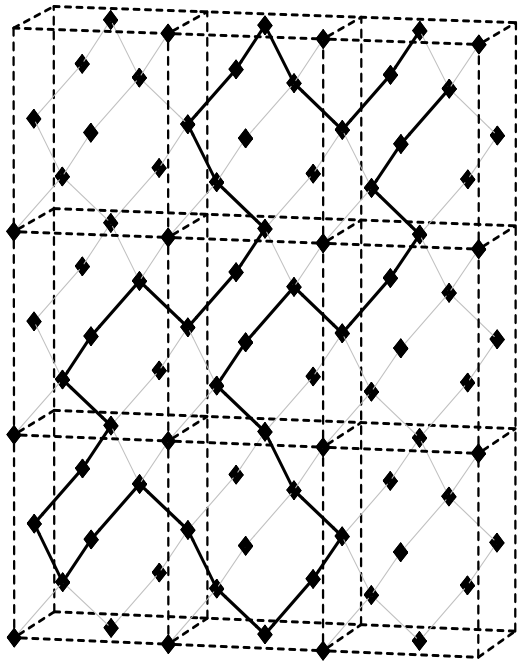

FIGURE 1.12: Illustration of a latticoid embedding into the diamond crystal lattice.

The bipartite and non-bipartite master approximation algorithms have performance guarantees for a class of lattices that includes most of the lattices commonly used in simple exact PSP models [27]: square and cubic lattices [18, 22, 39], diamond (carbon) lattices [40], face-centered-cubic lattice [14], and the 210 lattice used by Skolnick and Kolinkski [41]. Additionally, their analysis provides performance guarantees for a wide range of crystallographic lattices: Bravais lattices like the triclinic and triagonal lattices [38], the flourite lattice, 3D close packed lattices, the body centered cubic lattic and the hexagonal lattice. These results demonstrate that approximability is a general feature of HP models on crystal lattices.

1.4.5 Accessible Surface Area Lattice Model

Solvent accessible area (ASA) describes the surface area over which contact between protein and solvent can occur. The concept of the solvent accessible surface of a protein molecule was originally introduced by Lee and Richards [30] as a way of quantifying hydrophobic burial. Subsequently, ASA and similar measures have been integrated into a variety of empirical potentials for PSP. This potential is qualitatively different from the HP model in that it favors hydrophobic burial rather than hydrophobic-hydrophobic interactions.

We describe new performance guaranteed approximation algorithms for the ASA lattice model with a linear chain model on the triangular lattice. As with the HP model, this model considers protein sequences $s \in \{H, P\}^+$. On a lattice, the ASA for a protein conformation can be modelled by the number of unoccupied lattice points that are adjacent to hydrophobic amino acids. Since there exist sequences for which the ASA is zero (i.e. all

hydrophobics can be buried), it is not possible to develop an approximation algorithm that guarantees a multiplicative approximation ratio. Consequently, we treat this as a covering problem for the hydrophobics in a HP sequence.

Let $\overline{ASA}(s)$ refer to the number of covered hydrophobics in a conformation, which is the value we will attempt to maximize. If a sequence s has $N(s)$ hydrophobics, then $\overline{ASA}(s) \leq 4N(s)+2$ on the triangular lattice because each amino acid has four neighboring lattice points that are not covered by the chain itself (except for the endpoints). Let $N_{HP}(s)$ denote the number of H-P contacts in a conformation, and let $N_{HH}(s)$ denote the number of H-H contacts. Note that $\overline{ASA}(s) = N_{HP}(s) + 2N_{HH}(s)$, since a single hydrophobic-hydrophobic contact represents the fact that two hydrophobics are being covered. Now consider the conformation of a chain folded back on itself (a simple U-fold). All but 3 hydrophobics in this conformation are guaranteed to have two contacts. Consequently, $N_{HP}(s) + 2N_{HH}(s) \geq 2N(s) - 6$, so an algorithm that generates this conformation has a 1/2 approximation ratio. A similar analysis applies for a U-fold on the 2D square lattice, so an algorithm that generates that conformation has a 1/2 approximation ratio.

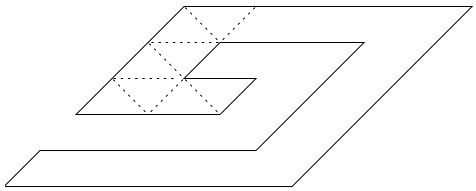

FIGURE 1.13: Illustration of conformations generated by the second approximation algorithm in the ASA model on a triangular lattice.

Now consider the conformation in Figure 1.13, which treats the protein as a circular conformation that is molded into a square shape. If the protein sequence has n amino acids, then approximately $n-4\sqrt{n}$ amino acids lie strictly within this conformation and are completely buried. Now consider the linear-time algorithm that shifts the protein sequence through the circular conformation to find the shift that minimizes the number of hydrophobics on the exterior of this conformation. The conformation of this shifted minimal sequence has at most $4N(s)/\sqrt{n}$ exposed hydrophobics. Thus we have $N_{HP}(s)+2N_{HH}(s) \geq 4N(s)-16N(s)/\sqrt{n}$, from which it follows that we have an approximation ratio of 1. This implies that asymptotically all but an $o(1)$ fraction of the hydrophobic amino acids are buried in this algorithm. Note that the conformation in Figure 1.13 can be embedded in the 2D square lattice. A similar analysis shows that an algorithm that generates this conformation has an approximation ratio of 1. Furthermore, this result naturally generalizes to the 3D cubic and FCC lattices, since you can create similarly compact structures for which the surface area is dominated by the volume.

1.5 Exact Methods

Solving PSP problems exactly is an important practical goal because the lowest-energy structure determines the biological functionality of a protein. Although PSP has been proven NP-hard for many different lattice models, this does not preclude the development of practical tools for many protein sequences. Since exhaustive enumeration is clearly not

practical even for relatively small protein sequences, several search techniques have been developed to solve PSP for simple lattices models. In each of these methods, the lattice structure is exploited to mathematically limit the search process.

1.5.1 Enumeration of Hydrophobic Cores

Yue and Dill [45] developed the first exact method for exactly finding globally optimal protein structures on HP lattice models. The surface area of the hydrophobic core is easier to estimate (given partial information about the final conformation) than the number of hydrophobic-hydrophobic contacts, and that the core surface area and the number of contacts are related one-to-one. Yue and Dill developed the constrained hydrophobic core construction (CHCC) algorithm, which enumerates all possible shapes of the region containing all hydrophobic amino acids for a sequence. This enumeration of the possible hydrophobic cores is done so that core shapes with a smaller surface area are enumerated before core shapes with a larger surface area. For every core shape, CHCC enumerates all positions of the monomers that fit into the given core shape. CHCC uses some conditions (or constraints) to reduce the size of the search tree.

The CHCC has been effectively applied to exactly solve PSP problems for HP sequences with up to 80 amino acids. Perhaps the greatest limitation of this method is that it is specifically tailored for the HP-model on the cubic lattice. Consequently, this basic algorithmic approach has not been effectively generalized to other simple lattice models for PSP.

1.5.2 Constraint Programming

Backofen et al. [3, 4, 5, 6, 7, 8] provide a declarative formulation of the HP lattice model, which is solved using constraint programming. Constraint programming is a relatively new programming technique that integrates a declarative definition of a problem (e.g. PROLOG) with an inherently concurrent programming paradigm, since all constraints are handled in parallel. The search strategy is not fixed in constraint programming, and systems like Oz [42] offer a flexible environment for defining a search strategy. Constraint programming offers a flexible framework for solving PSP on simple lattice models, and Backofen et al. have described declarative formulations for the HP models on the cubic and FCC lattices, as well as an extended HP model on the cubic lattice.

We illustrate the type of declarative formulation used for constraint programming to define feasible conformations in the cubic lattice. Consider variables X_i, Y_i and Z_i that indicate the position of the i-th amino acid in the lattice. Without loss of generality we can restrict the amino acids with the following constraint:

$$\forall i, \mathrm{X}_i \in [1 \ldots (2 \cdot \mathrm{length}(s))] \bigwedge \mathrm{Y}_i \in [1 \ldots (2 \cdot \mathrm{length}(s))] \bigwedge \mathrm{Z}_i \in [1 \ldots (2 \cdot \mathrm{length}(s))],$$

where length(s) is the length of the HP sequence. We clearly need to satisfy the constraint $\forall i \neq j, (X_i, Y_i, Z_i) \neq (X_j, Y_j, Z_j)$ in a feasible conformation. Additionally, amino acids must be consecutively placed on the lattice. We can enforce this constraint using variables Xdiff_i, Ydiff_i and Zdiff_i, which represent the difference of the x, y and z coordinates between amino acid i and $i+1$. The constraints

$$\begin{aligned} \forall i, \mathrm{Xdiff}_i &= |\mathrm{X}_i - \mathrm{X}_{i+1}| \\ \forall i, \mathrm{Ydiff}_i &= |\mathrm{Y}_i - \mathrm{Y}_{i+1}| \\ \forall i, \mathrm{Zdiff}_i &= |\mathrm{Z}_i - \mathrm{Z}_{i+1}| \end{aligned}$$

define the values of these variables, and the constraint $\forall i, 1 = \mathrm{Xdiff}_i + \mathrm{Ydiff}_i + \mathrm{Zdiff}_i$ ensures that the distance between consecutive amino acids is one.

Backofen et al. apply a search algorithm that is a combination of a branch-and-bound search together with a constrain-and-generate principle, which is common for constraint programming. The branching process selects a variable **var** to branch on and then creates two branches for some value **var**: (1) **var** =: **val**, and (2) **var** $\neq$: **val**. Subsequently, these branches are evaluated using a constraint programming system to evaluate the effected variables according to the constraints, which results in an association of smaller value ranges to some (or many) variables. Further, the search tree may be pruned when an inconsistent conformation is generated. The bounding calculation used in this search requires a problem-specific calculation, based on the feasible domain for a subproblem.

Backofen et al [3, 4, 5, 6, 7, 8] have evaluated constraint programming implementations for HP lattice models using the Oz language [42]. These methods have effectively solved problems of up to 200 amino acids (using pre-calculated hydrophobic cores) within a few seconds. Additionally, these tools have been used to enumerate optimal conformations for the HP cubic model, for which it appears to be more effective than the CHCC algorithm.

1.5.3 Integer Programming

A standard approach for finding exact solutions for hard optimization problems is to model them as integer programs and try to solve these programs to optimality using techniques from the field of integer programming such as branch and bound. Additionally, linear programming relaxations of integer programs often provide efficiently computable non-trivial upper bounds.

Several integer programming formulations have been developed for the PSP problem in the HP model [12, 21, 13]. We illustrate the type of linear constraints used for integer programming to define feasible conformations in the square lattice. Without loss of generality, we can restrict the conformations to lattice points $\mathcal{L} = \{1, 2, \ldots, n^2\}$, such that the coordinates are of the form:

$$y_p = \left\lfloor \frac{p-1}{n} \right\rfloor \text{ and } x_p = p - 1 - ny_p \text{ for } p \in \mathcal{L}.$$

Let $\mathcal{N}(p)$ denote the lattice points adjacent to a point p (whose distance is one away), and let v_{ip} be a binary decision variable that is one if the i-th amino acid is placed at point p on the lattice, and zero otherwise. Now every residue must be placed on a lattice point, which is enforced by the following constraint:

$$\sum_{p \in \mathcal{L}} v_{ip} = 1 \quad , i = 1, \ldots, n.$$

Similarly, each point cannot have more than one amino acid placed at it, which is enforced by the constraint:

$$\sum_{i=1}^{n} v_{ip} \leq 1 \quad , \forall p \in \mathcal{L}.$$

Finally, we can enforce the connectivity between consecutive amino acids with the following two constraints:

$$\begin{aligned} \sum_{q \in \mathcal{N}(p)} v_{i+1,q} &\geq v_{ip} \quad , i = 1, \ldots, n-1, \; p \in \mathcal{L} \\ \sum_{q \in \mathcal{N}(p)} v_{i-1,q} &\geq v_{ip} \quad , i = 2, \ldots, n, \; p \in \mathcal{L}. \end{aligned}$$

These constraints define a convex region that represents valid solutions if we relax the constraint that the v_{ip} are binary. This observation provides a mechanism for computing lower bounds on the minimum energy of a conformation with integer program formulations, for which the lower bound can be computed with linear programming methods.

Linear programming relaxations can provably provide bounds that are at least as strong as the simple combinatorial bound (1.3) and some IP formulations may strengthen this bound even further [12]. Although integer programming formulations have been used to compute such bounds, these formulations can have many variables, which may limit their application to large-scale problems. Additionally, it is not clear whether these integer programming formulations can be used to solve large-scale instances of the PSP problems exactly.

1.6 Conclusions

There are many ways that these analyses and methods for PSP problems can be improved. For example, no intractability analysis has been developed for the HP model on the triangular or FCC lattices. There is wide agreement that these lattices are more practically relevant for PSP because they do not impose the artificial parity found in the square and cubic lattices, so such an intractability analysis would be quite interesting. Similarly, exact methods have not been developed for models like the HP side chain model, which capture greater physical detail. We expect that studies of (near-) optimal conformations in this model would provide significant insight into PSP (e.g. by studying the degeneracy of the optimal solution in these problems).

Improving bounds on lattice models could fundamentally improve our assessment for approximation algorithms. For example, there are strings for which the best conformation on the 2D square lattice achieves only half of the upper bound in Equation (1.3) [33], so this bound is demonstrably weak. However, integer programming formulations may provide a general technique for improving these bounds for specific sequences. The bounds for the HP tangent spheres model might also be improved by generalizing the bound analysis of triangular and FCC lattices. In the triangular and FCC lattices, the bounds on the maximal number of contacts can be tightened by noting that "conflicts" occur between some hydrophobics and non-hydrophobics, thereby limiting the total number of hydrophobic-hydrophobic contacts. However, in 3D it is possible to have 12 spheres touching a given sphere without any pair of them being tangent, so the notion of a "conflict" needs to be generalized in this case to tighten simple upper bounds.

Researchers analyzing PSP in lattice models have increasingly considered detailed models and methods that can be applied to a variety of lattice models. This trend is motivated by the desire to provide robust mathematical insight into protein models that is generally independent of a particular lattice formulation. Analyses that achieve this goal provide greater insight into general PSP complexity, which is not bound by lattice constraints and for which precise empirical energy potentials are not known.

One interesting direction for the analysis of PSP is to consider methods that are tailored to biologically plausible amino acid sequences. Thus we need to develop complexity analyses like Heun's approximation algorithm that is tailored to protein-like sequences. For example, the possible intractability of PSP remains an open question if PSP is restricted in this manner.

Similarly, we expect that methods that can solve more detailed protein models will provide more insight into real protein structures. For example, side chain lattice models are clearly more representative of the structure of actual proteins than linear chain models. However,

the analysis of side chain models with variable-size side chains could more accurately capture the complexity of solving side chain packing problems. Additionally, this type of PSP formulation could capture the fact that the hydrophobicity of a side chain is related to its surface area. PSP with variable hydrophobicities has been briefly considered by Agarwala et al. [1], who consider protein structures as linear chains.

Finally, the connection between lattice models and off-lattice models needs to be developed further to more directly impact real-world PSP problems. Performance guaranteed algorithms for the FCC lattice can provide performance guarantees for closely related off-lattice protein models. This is a first step towards a more comprehensive analysis that uses lattice models to provide mathematical insight into off-lattice models. For example, we conjecture that lattice-based search methods like constraint programming can be effectively hybridized with optimizers for standard empirical energy potentials to perform a more effective global search of protein structures.

Acknowledgments

We thank Sorin Istrail for his collaborations on the ASA model. We also thank Edith Newman for her assistance in creating Figure 1.7. This work was performed in part at Sandia National Laboratories. Sandia is a multipurpose laboratory operated by Sandia Corporation, a Lockheed-Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000. This work was partially funded by the US Department of Energy's *Genomes to Life* program (`www.doegenomestolife.org`), under project "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling," (`www.genomes-to-life.org`).

References

[1] R. Agarwala, S. Batzogloa, V. Dančík, S. E. Decatur, S. Hannenhalli, M. Farach, S. Muthukrishnan, and S. Skiena. Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model. *J Comp Bio*, 4(3):276–296, 1997.

[2] J. Atkins and W. E. Hart. On the intractability of protein folding with a finite alphabet of amino acids. *Algorithmica*, 25:279–294, 1999.

[3] R. Backofen. Using constraint programming for lattice protein folding. In R. B. Altman, A. K. Dunker, L. Hunter, and T. E. Klein, editors, *Pacific Symposium on Biocomputing (PSB'98)*, volume 3, pages 387–398, 1998.

[4] R. Backofen. An upper bound for number of contacts in the HP-model on the Face-Centered-Cubic Lattice (FCC). In R. Giancarlo and D. Sankoff, editors, *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*, number 1848 in LNCS, pages 277–292, Montréal, Canada, 2000. Springer-Verlag, Berlin.

[5] R. Backofen. The protein structure prediction problem: A constraint optimisation approach using a new lower bound. *Constraints*, 6:223–255, 2001.

[6] R. Backofen and S. Will. Optimally compact finite sphere packings — hydrophobic cores in the FCC. In *Proc. of the 12th Annual Symposium on Combinatorial Pattern Matching (CPM2001)*, volume 2089 of *Lecture Notes in Computer Science*, Berlin, 2001. Springer–Verlag.

[7] R. Backofen and S. Will. A constraint-based approach to structure prediction for simplified protein models that outperforms other existing methods. In *Proceedings of the Ninetheen International Conference on Logic Programming (ICLP 2003)*, 2003. in press.

[8] R. Backofen, S. Will, and E. Bornberg-Bauer. Application of constraint programming techniques for structure prediction of lattice proteins with extended alphabets. *J. Bioinformatics*, 15(3):234–242, 1999.

[9] B. Berger and T. Leighton. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *J Comp Bio*, 5(1):27–40, 1998.

[10] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000. The PDB is at `http://www.rcsb.org/pdb/`.

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

[12] R. Carr, W. E. Hart, and A. Newman. Discrete optimization models for protein folding. Technical report, Sandia National Laboratories, 2003.

[13] V. Chandru, A. DattaSharma, and V. S. A. Kumar. The algorithmics of folding proteins on lattices. *Discrete Applied Mathematics*, 127(1):145–161, Apr 2003.

[14] D. G. Covell and R. L. Jernigan. *Biochemistry*, 29:3287, 1990.

[15] T. E. Creighton, editor. *Protein Folding*. W. H. Freeman and Company, 1993.

[16] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis. On the complexity of protein folding. *J Comp Bio*, 5(3), 1998.

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

[18] 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.

[19] C. J. Epstein, R. F. Goldberger, and C. B. Anfinsen. The genetic control of tertiary protein structure: Studies with model systems. In *Cold Spring Harbor Symposium on Quantitative Biology*, pages 439–449, 1963. Vol. 28.

[20] A. S. Fraenkel. Complexity of protein folding. *Bull. Math. Bio.*, 55(6):1199–1210, 1993.

[21] H. J. Greenberg, W. E. Hart, and G. Lancia. Opportunities for combinatorial optimization in computational biology. *INFORMS Journal of Computing*, 2003. (to appear).

[22] A. M. Gutin and E. I. Shakhnovich. Ground state of random copolymers and the discrete random energy model. *J. Chem. Phys.*, 98:8174–8177, 1993.

[23] W. E. Hart and S. Istrail. Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal. *Journal of Computational Biology*, 3(1):53–96, 1996.

[24] W. E. Hart and S. Istrail. Invariant patterns in crystal lattices: Implications for protein folding algorithms. In *Combinatorial Pattern Matching*, Lecture Notes in Computer Science 1075, pages 288–303, New York, 1996. Springer.

[25] W. E. Hart and S. Istrail. Lattice and off-lattice side chain models of protein folding: Linear time structure prediction better than 86% of optimal. *Journal of Computational Biology*, 4(3):241–259, 1997.

[26] W. E. Hart and S. Istrail. Robust proofs of NP-hardness for protein folding: General lattices and energy potentials. *Journal of Computational Biology*, 4(1):1–20, 1997.

[27] W. E. Hart and S. Istrail. Invariant patterns in crystal lattices: Implications for protein folding algorithms. *Journal of Universal Computer Science*, 6(6):560–579, 2000.

[28] V. Heun. Approximate protein folding in the HP side chain model on extended cubic lattices. *Discrete Applied Mathematics*, 127(1):163–177, 2003.

[29] 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.

[30] B. Lee and F. M. Richards. The interpretation of protein structures: Estimation of

static accessibility. *J Mol Biol*, 55:379–400, 1971.

[31] G. Mauri, A. Piccolboni, and G. Pavesi. Approximation algorithms for protein folding prediction. In *Proceedings of the 10th ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 945–946, Baltimore, 1999.

[32] A. Nayak, A. Sinclair, and U. Zwick. Spatial codes and the hardness of string folding problems. *J Comp Bio*, pages 13–36, 1999.

[33] A. Newman. A new algorithm for protein folding in the HP model. In *Proceedings of the 13th ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 876–884, San Francisco, Jan 2002.

[34] A. Newman and M. Ruhl. Combinatorial problems on strings with applications to protein folding. In *Proceedings of Latin American Theoretical Informatics (LATIN)*, Buenos Aires, 2004. To Appear.

[35] J. T. Ngo and J. Marks. Computational complexity of a problem in molecular structure prediction. *Protein Engineering*, 5(4):313–321, 1992.

[36] 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.

[37] M. Paterson and T. Przytycka. On the complexity of string folding. *Discrete Applied Mathematics*, 71:217–230, 1996.

[38] D. E. Sands. *Introduction to Crystallography.* Dover Publications, Inc., New York, 1975.

[39] E. I. Shakhnovich and A. M. Gutin. Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci.*, 90:7195–7199, 1993.

[40] A. Sikorski and J. Skolnick. Dynamice Monte Carlo simulations of globular protein folding/unfolding pathways. II. α-helical motifs. *J. Molecular Biology*, 212:819–836, July 1990.

[41] J. Skolnick and A. Kolinski. Simulations of the folding of a globular protien. *Science*, 250:1121–1125, 1990.

[42] G. Smolka. *The Oz programming model*, volume 1000 of *Lecture Notes in Computer Science*, pages 324–343. 1995.

[43] R. Unger and J. Moult. Finding the lowest free energy conformation of a protein is a NP-hard problem: Proof and implications. *Bull. Math. Bio.*, 55(6):1183–1198, 1993.

[44] W. F. van Gunsteren, P. K. Weiner, and A. J. Wilkinson, editors. *Computer Simulation of Biomolecular Systems.* ESCOM Science Publishers, 1993.

[45] K. Yue and K. A. Dill. Sequence-structure relationships in proteins and copolymers. *Phys. Rev. E*, 48(3):2267–2278, 1993.

References