Combinatorial Algorithms for Protein Folding in Lattice Models: A Survey of Mathematical Results

Sorin Istrail*
Center for Computational Molecular Biology
Department of Computer Science
Brown University
Providence, RI 02912

Fumei Lam†
Center for Computational Molecular Biology
Department of Computer Science
Brown University
Providence, RI 02912

Dedicated to Michael Waterman's 67th Birthday
June 17, 2009

"... a very nice step forward in the computerology of proteins." Ken Dill 1995[1]

Abstract

We present a comprehensive survey of combinatorial algorithms and theorems about lattice protein folding models obtained in the almost 15 years since the publication in 1995 of the first protein folding approximation algorithm with mathematically guaranteed error bounds [60]. The results presented here are mainly about the HP-protein folding model introduced by Ken Dill in 1985 [37]. The main topics of this survey include: approximation algorithms for linear-chain and side-chain lattice models, as well as off-lattice models, NP-completeness theorems about a variety of protein folding models, contact map structure of self-avoiding walks and HP-folds, combinatorics and algorithmics of side-chain models, bi-sphere packing and the Kepler conjecture, and the protein side-chain self-assembly conjecture. As an appealing bridge between the hybrid of continuous mathematics and discrete mathematics, a cornerstone of the mathematical difficulty of the protein folding problem, we show how work on 2D self-avoiding walks contact-map decomposition [56] can build upon the exact RNA contacts counting formula by Mike Waterman and collaborators [96] which lead to renewed hope for analytical closed-form approximations for statistical mechanics of protein folding in lattice models. We also include in this paper a few new results, research directions within reach of rigorous results, and a set of open problems that merit future exploration.

*corresponding author `sorin@cs.brown.edu`
†Present address: University of California Davis, Department of Computer Science, Davis, `lam@cs.brown.edu`

Contents

1 Introduction

> *"The subject of chaos is characterized by an abundance of quantitative data, an unending supply of beautiful pictures, and a shortage of rigorous theorems. Rigorous theorems are the best way to give a subject intellectual depth and precision. Until you can prove rigorous theorems, you do not fully understand the meaning of your concepts."* Freeman Dyson 2009[43]

> *"The most vitally characteristic fact about mathematics is, in my opinion, its quite peculiar relationship to the natural sciences ... In modern empirical sciences it has become more and more a major criterion of success whether they have become accessible to the mathematical method or to the near-mathematical methods of physics. Indeed, throughout the natural sciences an unbroken chain of successive pseudomorphoses, all of them pressing toward mathematics, and almost identified with the idea of scientific progress, has become more and more evident. Biology becomes increasingly pervaded by chemistry and physics, chemistry by experimental and theoretical physics, and physics by very mathematical forms of theoretical physics.*
> *There is a quite peculiar duplicity in the nature of mathematics. One has to realize this duplicity, to accept it, and to assimilate it into one's thinking on the subject. This double face is the face of mathematics, and I do not believe that any simplified, unitarian view of the thing is possible without sacrificing the essence."* John von Neumann 1947 [103]

We present here mathematically rigorous results about computational problems formulated in lattice models and a few results on off-lattice models. "Mathematical results" refers to the existence of mathematical proofs and this is the criterion for including such results in this survey. The theorems presented here concern algorithms for finding the lowest-energy conformation of lattice proteins as well as various other related combinatorial problems.

"[P]rotein folding is a fascinating cross-disciplinary field that attracts scientists with different backgrounds and scientific cultures. They bring to the protein folding field the models and the way of thinking that are accepted of their respective background fields. Such diversity of scientific cultures is a great virtue of the protein folding field, in which physics, chemistry, biology, and mathematics meet. It is important for our cross-disciplinary field to discuss with balance both strong points and limitations of different approaches" [98].

The bias of this survey is towards computer science and combinatorics contributions and the follow up literature after the first approximation algorithm for protein folding with guaranteed error bounds published in 1995 [60]. The lattice model discussed in detail here, together with a set of generalizations, is the HP-model of Ken Dill. This biophysical model is among the most studied model in the protein-folding literature [39, 41] and plays a unique role in research on combinatorial, computational complexity analysis, and algorithmic foundations of protein folding in lattice and off-lattice models. The main types of rigorous results included here are (1) approximation algorithms for finding the minimum-energy folds with mathematically guaranteed error bound, (2) computational complexity analysis by establishing proofs of NP-completeness for the problem of finding the minimum-energy folds in various models, and (3) estimates of the number of "native" contacts in the optimal fold of a given protein sequence. There are a number of great survey articles that cover topics both similar to this one [90, 39, 63, 42, 29] and complementary to [94, 87, 42, 41].

The rest of the paper is organized as follows. Section 2 presents protein folding models. Section 3 and 4 contain comprehensive surveys of the literature on approximation algorithms,

and respectively, NP-completeness theorems about protein folding models. Sections 5 and 6 present two promising research directions and existing mathematical theory that could inspire new developments towards rigorous algorithms. In Section 5 we present results on contact map decompositions of self-avoiding walks, connection with analytical formulas for counting them and approximations of partition functions, and contact map overlap algorithms. Section 6 is devoted to side-chain models, connections with the celebrated solution to the Kepler conjecture, and the bi-sphere packing problem. In Section 7 we discuss a set of problems that we believe have solutions within reach. We also include a few new results, reflections on combinatorial strategies and approaches to the problems discussed, and a number of open problems whose solutions will advance our understanding of the field.

1.1 The Computational Protein Folding Problem

> *"The protein folding problem is three different problems: the folding code – the thermodynamic question of how a native structure results from the interatomic forces acting on an amino acid sequence; protein structure prediction – the computational problem of how to predict the native structure of a protein from its amino acid sequence; and the folding speed (Levinthal's paradox) – the kinetic question of how a protein can fold so fast ... Current knowledge of the folding codes is sufficient to guide the successful design of new proteins and new materials. Current computer algorithms are now predicting the native structures of small simple proteins remarkable accurately, contributing to drug discovery and proteomics. Even once intractable Levinthal puzzle now seems to have a very simple answer ..."* Ken Dill 2007 [40]

The protein folding problem is in fact a collection of fundamental problems focused on the questions, "What is the folding code?" and "What is the folding mechanism?" [41] and " ... the second, more visible to the public, side of the 'holy grail' of protein folding – prediction of protein conformation " [98]. The "folding code" concerns how the "tertiary structure and folding pathway of a protein are encoded in its amino acid sequence...[it] is not predominantly localized in short windows of the amino acid sequence ... [it] resigns mainly in global patterns of interactions, which are nonlocal, and arise from the arrangements of polar and non-polar monomers in the sequence" [39].

In this survey, we highlight research directions that attempt to relate the physical process of protein folding and the informal computational paradoxes relevant to folding models to the rigorous analysis of associated mathematical problems and insight about folding algorithms. Here are a few such themes.

Biophysics and algorithmics of folding. Mathematical proofs could in some measure provide the validity of a biophysical *modus operandi* or conjecture. For example, the "balancing points" existing in every protein sequence, on which approximation algorithms with guaranteed error bound are based, as presented in Section 3.1, correspond in spirit to the Zipper and Assembly hypothesis [41]. Similarly, it is common thinking that lattices are viewed interchangeably as ways of discretizing the 3D space. The "master approximation algorithms theorem" in section 3.1 proves a mathematical result with a similar "universal" flavor. It is also interesting to note that several proofs of NP-completeness, as presented in Section 4, have as a first step the "multistring"-folding version of the single-string folding problem. The multistring models have clear similarities with protein misfolding and aggregation, bringing into view the continuum between folding and misfolding as far as folding combinatorics goes.

"The failure of protein-folding processes, both within cells (in vivo) and within test tubes or industrial vats (in vitro), causes serious difficulties both for biomedical research and for biotechnology industry. Protein chains that fail to fold properly aggregate into an insoluble and inactive state... There is increased recognition that some human diseases are associated with aberrations or defects in protein chain folding. These include Alzheimer's and Huntington's and cystic fibrosis." Jonathan King 2002 [78]

Unavailability of foundational mathematical results. At the heart of the mathematical difficulty of computational protein folding is the hybrid between the geometry, rooted in continuous mathematics, and the combinatorics of folds, an essentially discrete mathematics theme. In addition, the discrete mathematics component involves the notoriously difficult concept of self-avoiding walk, a.k.a. excluded volume, which escapes even the classifying power of NP-completeness (#P-completeness). Optimization problems with such multi-criteria optimization of continuous-discrete type have basically no known foundational mathematical theory.

Fast approximation algorithms and the hydrophobic collapse. Section 3.1 presents a number of approximation algorithms for various lattice and off-lattice models. It is interesting that almost all such algorithms are very fast (linear time in the size of the protein sequence) and produce near-optimal folds. Even more, the master approximation algorithm 3.1 uses a formal concept of universality. This "universality" across models resembles the hydrophobic collapse phenomenon, postulated as a major driver for folding in globular proteins. In this sense, these models capture that aspect of folding well and there is a mathematical theorem capturing it.

The Kepler Conjecture and the "densest off-lattice is on-lattice" phenomenon. The history and the recent celebrated solution of the almost 400-year-old Kepler Conjecture highlight the notorious mathematical difficulty of packing problems with relevance for protein folding, as we discuss in Section 6. The conjecture formulated by Johannes Kepler in 1611 [77] states that the highest packing density of identical spheres is obtained by arranging them as nodes in a face-centered-cubic (FCC) lattice. Gauss showed in 1831 that, of all lattice arrangements, the FCC lattice is indeed the one that provides a packing of the entire 3D space with the highest density. Only at the end of the last century, in 1997, was the solution to the general problem including off-lattice arrangements obtained by Thomas Hales [59], showing that the densest off-lattice arrangement is again the FCC lattice. We discuss in Section 6 a conjecture analogous to the Kepler conjecture involving bi-spheres, the bi-sphere packing problem, and connect it with what we propose as the Protein Side-Chain Self-Assembly conjecture. The on-lattice-off-lattice proofs exemplify the enormous difficulty of such mathematical arguments highly relevant to folding combinatorics as well. It is this type of difficulty that is involved in extending the type of mathematical results on approximation algorithms presented in this survey for the HP-model to other leading biophysical lattice and off-lattice models.

2 Protein Folding: Models

"Understanding the mechanism of protein folding is often called the "second half" of genetics. Computational approaches have been instrumental in the efforts. Simplified models have been applied to understand the physical principles governing the folding processes and will continue to play important roles in the endeavor." Peter Kollman 2001[42]

"We must emphasize a statement which I am sure you have heard before, but which must be repeated again and again. It is that the sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes the observed phenomena. Furthermore, it must satisfy certain esthetic criteria, that is, in relation to how much it describes, it must be rather simple. Since one cannot tell exactly how 'simple' simple is ... Simplicity is largely a matter of historical background, of previous conditioning, of antecedents, of customary procedures, and it is very much a function of what is explained by it." John von Neumann 1955 [104]

2.1 Assumptions and Caveats

"[W]e take as our premise that proteins are chain molecules that have specific monomer sequences and are driven to fold mainly by nonlocal interactions subject to steric constraints. There is currently no accurate analytical theory that can account for chain connectivity, excluded volume in the compact states, and specific sequences of monomer units. Simple exact models have been developed to explore such properties." [39]

Protein folding models provide a mathematical formulation of the protein folding process, abstracting away components of atomic detail as well as making choices on what the model will include. Modeling is a choice. One could simplify, as is usually done, by considering only some forces to include in the model, but one can also introduce generality, such as considering all the HP-sequences as protein sequences. It is this simplify-and-generalize power of modeling that, as far as the computerology of proteins is concerned, is the source of some monotonicity fallacies we discuss next.

The "More-is-Harder" Monotonicity Fallacy Argument. By 1995, when the first approximation algorithms were published [60], the computer-science notion of "time" apparently did not enter essentially, as far as we know, into results on computational methods in protein folding, except that it was recognized as an "exponential problem" [94]. The use of the concept of polynomial time algorithm as an aim for tractable computational methods was just emerging in the protein folding research [46, 58].

In 1992 and 1993 the first papers containing proofs of NP-completeness for protein folding models were published. They were established for ad-hoc models introduced basically for the purpose of the NP-completeness proof.

It is natural to think that if we cannot solve a problem in a simple model than the problem could only be harder in a more complex model. Although this is intuitively true in informal arguments, it is a source of misguided arguments. We comment next about such a monotonicity fallacy both related to algorithms and to NP-completeness.

As far as algorithms for folding are concerned, the literature before 1995 contained mainly two types of computational methods: *exhaustive enumeration* for protein models, which applied only to relatively short protein sequences, and *stochastic methods* (simulations or sampling methods). For these types of methods a *monotonicity property* is certainly true: More-is-Harder. That is, if a "simple" model is difficult to solve in terms of computational resources, the same problems for a more "complex" model, e.g., involving more atomic detail, would require more computational resources. This is clearly true for exhaustive enumeration algorithms, and it is only informally true for stochastic methods. The following is a fallacy: "computational difficulty with methods in simple models has as a logical consequence difficulty in complex models." From the computerology point of view, in a more complex model, the problem under study may become more tractable due to the

extra constrains imposed. For example, in the case of approximation algorithms, the best approximation ratios to date for the side-chain HP-model are superior to the linear-chain HP-models (Section 3.1).

The same type of monotonicity property was asserted incorrectly in proofs of NP-completeness. If a class of instances of a computational problem is NP-complete, a larger class containing those, that is, a *generalization*, has the NP-completeness monotonically preserved. However, the fallacy comes in place when both *generalizations* and *simplifications* are considered. One such fallacious argument is as follows: Step 1: one introduces a protein folding model and shows that protein folding is NP-hard; Step 2: "Nature" solves protein folding very fast; Step 3: therefore, "Nature" solves NP-hard problems fast (in polynomial time).

2.2 Lattice Models

> *"It seems remarkable that so simple a model based on time averaged forces can account for the stability and folding of a molecule as complicated as a protein. Looking at known protein conformations closely, one is struck by the precise geometry of the interatomic contacts that stabilise the molecule: all possible interior hydrogen bonds are well formed, and many of the nonpolar side chains interlock to form a close packed interior. ... [T]he forces responsible for this precise geometry ... cause the chain to fold into the approximate shape rapidly and without having to pass through many local minima ... Although calculating the energy of the all atom molecule would be time consuming, one would have the great advantage of starting close to the right conformation ... The general concept of using a simple model ... when the detailed forces are too complicated has many potential applications ... Such a hierarchical approach might eventually lead to an understanding and simulation of very complicated biological assembly processes. "* Michael Levitt and Arieh Warshel 1975 [82]

> *"[F]olding is an intrinsically statistical phenomenon and no conclusion can be derived from a single folding or unfolding trajectory. ... Lattice and other simplified analytical models are the statistical mechanician's contribution to the protein folding ... their intimate connection with statistical mechanics ... is very important as it often allows us to compare simulation with statistical-mechanical analytical theories."* Eugene Shakhnovich 1996 [98]

We follow the classification of lattice models from Duan and Kollman [42]. The pioneering work of Levitt and Warshel in the 1970s creating the first detailed energy-minimization lattice model for studying the folding of BPTI marked "the beginning of the physically based models in the study of protein folding ... the fact that most current structure prediction methods use a similar representation to that of Levitt and Warshel is a strong testament of the power of such an approach" [82]. Lattice models are of two types. The first type is "designed to understand the *basic physics* governing the protein folding process," while the second type aims at "*realistic folding* of real proteins and are therefore parameterized using real proteins as templates by statistical sampling of the available structures and are often referred to as statistical potentials." In the first basic physics category, the major models that provided deep insights into the physical principles of folding are Go (simplicity) [54], Wolynes (funnel-like energy landscape) [110], Dill (hydrophobic interactions, hydrophobic-hydrophilic pattern) [39, 38], Shakhnovich (statistical mechanics) [95], Karplus (diffusion-collision) [111]. In the second realistic folding category, the leading models are

due to Skolnick [99], Miyazawa and Jernigan [84], Crippen [36], Eisenberg [22], Sippl [66], Scheraga [83].

The 2D square lattice and the 3D cubic lattice are the most thoroughly studied lattices and consequently have extensive literature on exact computational methods, approximation algorithms, and complexity results.

In three dimensions, a lattice of major importance is the face-centered-cubic (FCC) lattice. It has been shown that the neighborhood of amino acids in proteins closely resembles an FCC lattice, providing evidence for the importance of the FCC lattice in modeling protein folds [18, 19]. Furthermore, the *kissing number* of a sphere in 3D space is known to be 12, the same as the degree of each vertex in the FCC structure [20, 70]. Therefore, the number of degrees of freedom for placing adjacent spheres in three dimensions is achieved by the vertices of the FCC lattice. This is intimately tied to Kepler's conjecture, recently proved by Thomas Hales, which states that the face-centered-cubic lattice is the densest packing of identical spheres in three dimensions [33, 100]. The face-centered-cubic lattice therefore provides the densest possible hydrophobic core for any lattice-based protein folding model. We discuss in Section 6 the connection between the much celebrated solution of the Kepler Conjecture and methods developed there that have relevance to protein folding packing problems.

2.2.1 The HP-Model

In 1985, Ken Dill proposed the hydrophobic-hydrophilic (HP) model, which has been subjected to a huge amount of literature due to its fundamental role in protein folding modeling [17, 37, 76, 81]. The model captures the fact that native protein folds tend to form very compact cores driven by dominant hydrophobic interactions [39]. Each amino acid is classified either as hydrophobic (H) or hydrophilic (P) and two hydrophobic amino acids are said to be in *contact* if they are adjacent in the fold but nonadjacent in the primary sequence. Since the goal is the formation of highly compact hydrophobic cores, the optimization function is to maximize the number of contacts between hydrophobic atoms (H-H contacts). To phrase the problem as an energy-minimization problem, the energy function is the negative of the number of hydrophobic contacts of the fold. Two examples of protein folds in the linear-chain and side-chain lattice HP-models are presented in Figures 1 and 4.

Models represent the protein sequence as a linear chain, perhaps with explicit side-chains branching from the linear backbone. In *lattice models*, a fold of a protein sequence is defined by placing the amino acids on lattice nodes and the protein chain as a self-avoiding path on the lattice; in *off-lattice models*, the placement of the protein is in 3D space, with the only restriction being the self-avoidance of the backbone and of the branching side-chains [39].

3 Protein Folding: Approximation Algorithms

> *"The central question addressed in this review is this: Is there some clever algorithm, yet to be invented, that can find the global minimum of a protein's potential-energy function reliably and reasonably quickly? Or is there something intrinsic to the problem that prevents such a solution from existing? ... Is there an approximation algorithm for global potential-energy minimization? ... To our knowledge, the possible existence of an approximation algorithm for protein structure prediction has not been addressed ... Such an approximation algorithm might be of significant practical use in protein-structure prediction, because exactness is not a central issue."* Martin Karplus 1994 [90]

"It is the mark of an instructed mind to rest satisfied with the degree of precision which the nature of the subjects permits and not seek an exactness where only an approximation of the truth is possible." Aristotle 319 BC [8]

The two types of combinatorial methods with rigorous mathematical results are *approximation algorithms* and *combinatorial optimization algorithms.*

The two classes of methods differ in their time complexity, whether they are exact or approximations, and whether they apply with the claimed performance to the entire class of protein sequences or a restricted subclass class of protein sequences.

An approximation algorithm for protein folding in an HP-model is a *polynomial-time* algorithm that for *every* protein sequence outputs a fold of that protein whose number of contacts is provably *near optimal.* A combinatorial optimization algorithm for protein folding in an HP-model is an *exponential* algorithm that for *some* protein sequences outputs a fold of that protein whose number of contacts is provable *optimal.* More critically, although approximation algorithms may be very fast and apply to all sequences, they could have weak approximation ratios and therefore may not be close to optimal (e.g., 0.5% of optimal means a fold with at least half of the number of optimal contacts). The best approximation algorithms to date have approximation ratio 86% of optimal [64]. On the other hand, combinatorial optimization algorithms, despite being exponential, find exact optimal solutions; however, only for small to moderate size proteins (about 50 amino acids), and even for a given size not all problems can be solved exactly and it is not clear in general which ones can be solved. Such methods were able to find optimal solutions for proteins of size 100 [14]. *Heuristics* are a third category of combinatorial methods. This is an area with a large amount of literature, including a variety of methods both deterministic and stochastic. These methods are not presented here. We present a comprehensive literature survey of approximation algorithms, while we only cite a few of the most successful results in the combinatorial optimization literature.

3.1 Approximation Algorithms: Fast Folding Provably Close to Optimal

In this section, we discuss approximation algorithms for HP-models. The results apply to lattice models, for linear-chain and side-chain HP-models, as well as generalizations to an off-lattice model, the Tangent Spheres HP-model.

When an optimization problem is showed to be computationally intractable, i.e., NP-complete, the next avenue to consider is the existence of algorithms that could compute solutions close to optimal. These are called Approximation Algorithms. They obtain provably near-optimal solutions of problems for which exact optimization is proved NP-complete. To quantify "closeness to optimal" one uses the following terminology. An α-approximation algorithm for a problem is a polynomial-time algorithm that outputs a solution of ratio at most α from the optimal solution. For a minimization problem, the output is at most α times the value of the optimal solution and for a maximization problem, the output is at least $\frac{1}{\alpha}$ times the value of the optimal solution. The value α is called the *approximation ratio* or *approximation guarantee* for the problem.

For protein folding in HP-models, the optimization problem is defined as follows. Given an HP-model, and a protein sequence over the binary alphabet of hydrophobic-hydrophilic amino acids, find the protein fold in the model that has the maximum number of contacts. This optimization problem is indeed NP-complete in many HP-models as shown in Section 4. However, a collection of approximation algorithms exist for a variety of HP-models.

We present next the ideas behind the first such algorithms and the gallery of approximation algorithms to date for HP-models and generalizations.

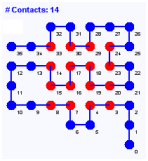

Figure 1: Optimal fold of a protein of length 36

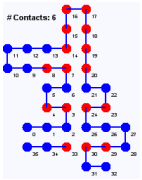

Figure 2: A fold constructed by the Hart-Istrail algorithm

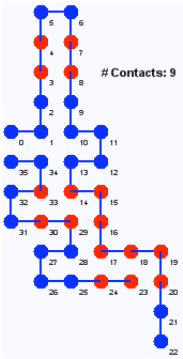

Figure 3: A fold constructed by the Newman algorithm

3.1.1 Hart-Istrail Algorithms

Let us consider the maximization problem with objective function C, maximizing the number of contacts of a fold.

The first step in the design of an approximation algorithm for this maximization problem is to find an upper bound on the optimal value of the objective function that can be computed in polynomial time. This upper bound for our problem would provide, for a given protein sequence, an estimate of the optimal number of contacts of the sequence in its "native" fold. For example, in the 2D square lattice, since each H can make at most two contacts, such an upper bound is two times the minimum number of H's that are even or odd. Finding good upper bounds, i.e., upper bounds that are as close to the optimal solution as possible, is crucial in developing algorithms with good approximation ratios.

We describe the main ideas of the first rigorous approximation algorithms in the literature with provable approximation guarantees for the protein-folding problem in the HP model on 2D and 3D cubic lattices, established by Hart and Istrail [60]. Consider a protein sequence S of hydrophobic (H) and hydrophilic (P) amino acids (or residues) and label the amino acids in the sequence in order by $1, 2, \ldots n$. A hydrophobic residue is said to be an *odd* hydrophobic if it is labeled by an odd number and an *even* hydrophobic otherwise. Let $\mathcal{O}(S)$ denote the number of odd hydrophobics in S and let $\mathcal{E}(S)$ denote the number of even hydrophobics in S. Let

$$\mathcal{C}_{2D}(S) = 2\min\{\mathcal{O}(S), \mathcal{E}(S)\}$$

The Hart-Istrail algorithms for protein folding in the 2D cubic lattice HP model on a 2D cubic lattice use the following estimate of the optimal number of contacts: for every fold $F(S)$ for the sequence S we have:

$$\#Contacts(F(S)) \leq \mathcal{C}_{2D}(S)$$

The upper bound $\mathcal{C}_{2D}(S)$ depends crucially on properties of the underlying two-dimensional square lattice, in particular that an $H-H$ contact in a fold can be formed only if the amino acids are in positions of different parity in the sequence.

The Hart-Istrail approximation algorithm proceeds as follows. The intuition is based on a *balancing point* (or a turning point) that exists in every protein sequence. Given a position p of an amino acid of the sequence S, let $L_S(p)$ denote the amino acids in S to the left of p and $R_S(p)$ denote the amino acids in S to the right of p. Find a position p such that at least half of the even hydrophobics fall on one side of p and at least half of the odd hydrophobics fall on the other side of p. Such a position exists for the following reason. Choose p such that $L_S(p)$ and $R_S(p)$ each contain exactly half the even ones; then either $L_S(p)$ or $R_S(p)$ contain at least half of the odd ones.

The approximation algorithm finds a fold that matches all the even hydrophobics on one side of p with all the odd hydrophobics on the other side of p by forming loops of all of the intermediary amino acids. By the choice of p and the construction of the fold, the number of contacts in the resulting fold is at least $\frac{1}{2}\min\{\mathcal{E}(S), \mathcal{O}(S)\}$. Therefore the approximation ratio achieved by the algorithm is

$$\frac{\frac{1}{2}\min\{\mathcal{E}(S), \mathcal{O}(S)\}}{2\min\{\mathcal{E}(S), \mathcal{O}(S)\}} = \frac{1}{4}.$$

Figure 2 shows the fold obtained by repeating the algorithm over all possible balancing points p in this algorithm.

Newman has given examples of sequences for which $\mathcal{C}_{2D}(S)$ is within a factor of 2 of the optimal number of contacts [88]. Therefore any approximation algorithm (which by definition works on all protein sequences) whose guaranteed close-to-optimal bound is based on the

$$\#Contacts(F(S)) \leq \mathcal{C}_{2D}(S)$$

inequality cannot achieve approximation ratio strictly greater than $\frac{1}{2}$.

For the 3D cubic lattice HP-model, Hart and Istrail gave a $\frac{3}{8}$-approximation for the protein-folding problem. For the 3D cubic lattice, it is easy to see that we can define

$$\mathcal{C}_{3D}(S) = 4(\min\{\mathcal{O}(S), \mathcal{E}(S)\}) + 2$$

and for every fold $F(S)$ for the sequence S we have:

$$\#Contacts(F(S)) \leq \mathcal{C}_{3D}(S)$$

Let $k = \frac{\mathcal{O}(S)}{2}$ and consider the position p such that the left and right sides $L_S(p)$ and $R_S(p)$ contain at least k odd and even hydrophobics, respectively. The next step is to divide $L_S(p)$ into segments with $\sqrt{k}$ odd hydrophobics and divide $R_S(p)$ into segments with $\sqrt{k}$ even hydrophobics. The Hart-Istrail 2D folding algorithm is then repeatedly applied $\sqrt{k}$ times in adjacent (x, y) planes. The idea is that each of the odd Hs in $L_S(p)$ has three contacts: one in the (x, y) plane, one with the plane above, and one with the plane below. In particular, three contacts are made for $\frac{\mathcal{O}(S)}{2} - o(\sqrt{\mathcal{O}(S)})$ odd Hs. This results in an algorithm with approximation guarantee of $\frac{3}{8}$ of optimal.

3.1.2 Newman Algorithm

In [88], Newman developed an improved approximation algorithm with approximation ratio $\frac{1}{3}$ for the 2D square lattice. While the upper bound used for analysis of the approximation algorithms is the same $\mathcal{C}_{2D}(S)$, the Newman algorithm uses additional bends in the sequence that improve the number of contacts between the two sides of the position p. Figure 3 shows the fold obtained by repeating the algorithm over all possible turning points p in this algorithm.

3.1.3 Linear-Chain Lattice, Side-Chain Lattice and Off-Lattice Models

Linear-Chain Models. The Hart-Istrail and Newman algorithms are approximation algorithms for linear-chain protein-folding models. A comprehensive gallery of approximation algorithms for linear-chain lattice, side-chain lattice, and off-lattice is presented in Section 3.2.

The existence of approximation algorithms for HP-models across lattices is a universal phenomenon that is formalized in the following theorem. In [62], Hart and Istrail design a *master approximation algorithm*, a general method for protein folding on the HP model that applies to a large class of lattice models; see Figure 5, for the "universal" type of sublattice structure responsible for the construction. These algorithms apply to most well-studied lattices in the literature, including the two-dimensional square lattice, the three-dimensional cubic lattice, the diamond lattice, the Bravais lattice, the FCC lattice, and other crystallographic lattices. The approximation algorithms in their paper correspond to two different algorithms that must be applied separately to the cases of bipartite and nonbipartite lattices. Both algorithms follow the structure of the approximation algorithms in Section 3.1.1 by first selecting a point to balance the number of hydrophobics in each side, and then forming a backbone or core with guidance from the formation of hydrophobic edge contacts. These approximation algorithms can also be generalized to give approximation guarantees for folding in off-lattice protein models [64].

Theorem 3.1 *There is a linear time master approximation algorithm universal to all HP-models across lattices.*

Side-Chain Models. To increase the accuracy of protein prediction methods, it is desirable that extended models take into account the structure of the protein as a backbone formed by a set of successive peptide bonds, together with attached side chains [23]. Ngo, Marks, and Karplus ask for the computational complexity of protein folding in models that include side chains [90].

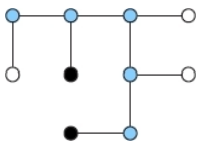

Figure 4: A fold of a protein in the 2D square side-chain HP-model making one contact

The side-chain lattice models analyzed represent the folding of proteins as "branched combs." In the side-chain model, the backbone of the protein is represented by a linear sequence of backbone nodes (as in the HP-model, except that these nodes are not labeled with amino acids), and connected to each backbone node is a side chain (an edge with one end a backbone node and the other end representing the amino acid) representing an

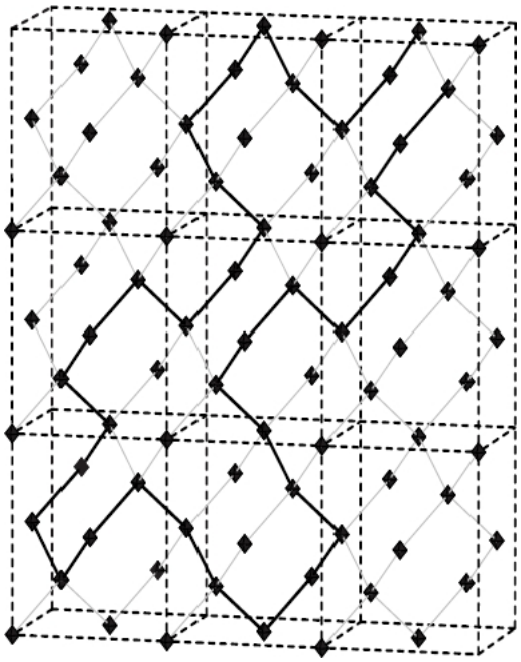

Figure 5: The "universal" sublattice structure responsible for the construction of the master approximation algorithm.

amino acid (labeled either hydrophobic or hydrophilic) (Figures 4 and 16). A conformation of the protein is an embedding in a "self-avoiding manner" of the backbone path into the lattice with side-chain edges mapped to adjacent lattice edges such that no lattice point is occupied by more than one backbone node or side-chain node. As before, the energy of a conformation is the number of hydrophobic-hydrophilic contacts between amino acids.

The algorithms in Hart and Istrail [64] were the first to provide approximation guarantees for the problem of folding on the side-chain model. Their results apply to 2D square, 3D cubic lattices, and FCC lattices. For cubic lattices, their algorithm proceeds by decomposing the sequence into a sequence of *blocks* that satisfy certain constraints on the parity and structure of hydrophobic amino acids. The structure in these blocks is used to prove upper bounds on the number of possible contacts in any conformation. The algorithms in [64] follow the approach of approximation algorithms in the HP model without side chains, first choosing a balancing or turning point of the sequence that balances the number of hydrophobics on either side of the turning point, and then matching up the hydrophobics from each partition to form a hydrophobic core. The pattern of the fold in the algorithm depends on the hydrophobic-hydrophilic pattern of the sequence. The resulting approximation algorithms have approximation ratio 1/12 in the 2D square lattice and 2/5 in the 3D cubic lattice.

In the side-chain model on the FCC lattice, the sequence is partitioned into eight subsequences with approximately equal numbers of hydrophobics in each part. Each of these parts is then placed in a single column of the FCC lattice forming the hydrophobic core. The hydrophilics are looped in the fold to form disjoint sets of columns, using different layers of the FCC lattice for each loop. We refer the reader to [64] for illustrations of the hydrophobic core construction and looping structure. This algorithm differs from all the

previous approximation algorithms described in that it uses the concept of balance point in a recursive fashion to separate the sequence into parts each making a fraction of the number of optimal contacts. Both the partitioning and looping can be done in linear time, giving rise to a linear-time approximation algorithm with approximation ratio 31/36 ($> 86\%$). This ratio remains to date the best ratio for an approximation algorithm in any 3D HP-models. In addition to this algorithmic development, it is shown how prior algorithms for the cubic lattice in 2D (with and without side chains) can be used to develop algorithms for the side-chain model in the 3D cubic lattice.

Off-Lattice Models. An important question to address in studying lattice models is whether or not the algorithms for these models can be generalized to algorithms for off-lattice models. In 1997, Hart and Istrail introduced an off-lattice model called the Tangent Spheres side-chain HP-model (HP-TSSC) [64]. In this model, adjacent backbone and side-chain molecules are represented by identical spheres in 3D space that are tangent. Side chains are labeled hydrophobic or hydrophilic and the energy of a conformation is the number of hydrophobic-hydrophobic tangent spheres.

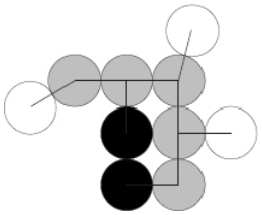

Figure 6: A fold for a protein in the 2D Tangent Spheres HP-model

Hart and Istrail demonstrate that algorithms with provable guarantees for protein folding in lattice models lead to algorithms with provable guarantees for the folding problem in off-lattice models, illustrating the transformation of approximation guarantees from lattice models to off-lattice models [64]. In particular, the approximation algorithm for the side-chain model in the FCC lattice, as described in Section 3.2.2, can be cast in the off-lattice framework and provides an off-lattice performance guarantee that is close to optimal (86% of optimal). The proof of upper bounds for the total number of possible contacts in the tangent spheres model is closely related to the number of neighbors in the FCC lattice, as follows from the well-known *kissing number* in 3D space [59].

Since the number of neighbors in the FCC lattice and the maximum number of possible neighbors in space is the same, this suggests that algorithms on the FCC lattice have good off-lattice performance. This is an example of using the mathematical principles of packings in 3D space to constrain the model in order to make the resulting problems more tractable. Similarly, a general method is given in [64] to cast the algorithms established by lattice models to off-lattice frameworks to achieve rigorous approximation guarantees. For example, the approximation algorithms of Agarwala et al. in [5] can be cast in the off-lattice framework to achieve approximation ratio of 54.5%. The following figures illustrate folds in the 2D hexagonal (triangular) lattice, 3D linear-chain cubic lattice, and 3D side-chain FCC lattice.

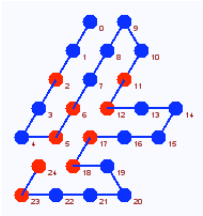

Figure 7: A fold in the 2D hexagonal (triangular) lattice for a protein of length 25

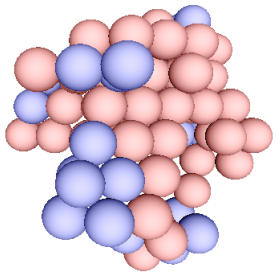

Figure 8: A fold of a protein of length 85 in the 3D Face Centered Cubic lattice

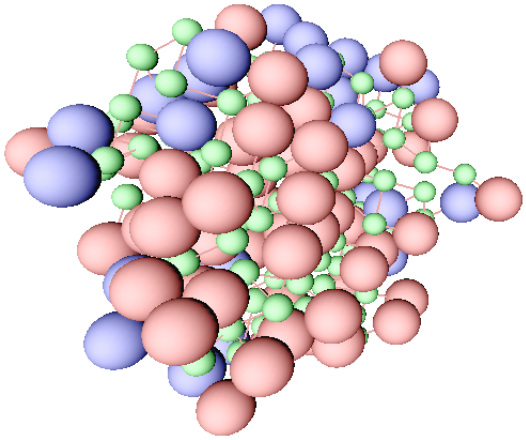

Figure 9: A fold in the side-chain HP-model on the Face Centered Cubic lattice

3.2 The Gallery of Approximation Algorithms For Lattice and Off-Lattice Models

We now present the theorems concerning the approximation algorithms. They all resemble the first such approximation algorithm in that almost all are linear-time algorithms and use a combination of global and local folding rules, although they differ in the underlying combinatorics used to achieve the close-to-optimal folding.

3.2.1 Approximation Algorithms for Linear-Chains Models

1. Triangular Lattice:

   Theorem 3.2 (Agarwala-Batzoglou-Dancik-Decatur-Farach-Hannenhalli-Skiena 1997) *There is a linear-time approximation algorithm that folds an arbitrary HP-protein sequence in the 2D triangular HP-model within $\frac{6}{11}$ (0.55) of optimal.*

   Theorem 3.3 (Agarwala-Batzoglou-Dancik-Decatur-Farach-Hannenhalli-Skiena 1997) *There is a linear-time approximation algorithm that folds an arbitrary HP-protein sequence in the 3D triangular HP-model within $\frac{44}{75}$ (0.59) of optimal.*

   Theorem 3.4 (Batzoglou-Decatur 1996) *There is a linear-time approximation algorithm that folds an arbitrary HP-protein sequence in the 2D triangular HP-model within $\frac{1}{2}$ (0.5) of optimal.*

   Theorem 3.5 (Batzoglou-Decatur 1996) *There is a linear-time approximation algorithm that folds an arbitrary HP-protein sequence in the 3D triangular HP-model within $\frac{3}{5}$ (0.6) of optimal.*

2. 2D Square Lattice:

   Theorem 3.6 (Hart-Istrail 1995) *There is a linear-time approximation algorithm that folds an arbitrary HP-protein sequence in the 2D square HP-model within $\frac{1}{4}$ (0.25) of optimal.*

   Theorem 3.7 (Hart-Istrail 1995) *For every protein sequence S in the linear chain HP-model, consider the following:*

   (a) $OPT_{2D}(S)$ = the maximal number of contacts of any fold of S on the 2D square lattice; $OPT_{3D}(S)$ = the maximal number of contacts of any fold of S on the 3D cubic lattice;

   (b)

   $$\mathcal{C}_{2D}(S) = 2\min\{\mathcal{O}(S), \mathcal{E}(S)\}$$
   $$\mathcal{C}_{3D}(S) = 4(\min\{\mathcal{O}(S), \mathcal{E}(S)\}) + 2$$

   Then

   (a) $OPT_{2D}(s) \leq \mathcal{C}_{2D}(S)$
   (b) $OPT_{3D}(s) \leq \mathcal{C}_{3D}(S)$

Therefore, every 2D algorithm that constructs folds achieving a fraction of α of the $\mathcal{C}_{2D}(S)$ contacts is an approximation algorithm with ratio α; it is then guaranteed to achieve at least α of the optimal number of contacts. Similarly for the 3D case.

Theorem 3.8 (Hart-Istrail 1995) *There is a linear-time approximation algorithm that folds an arbitrary HP-protein sequence in the 3D cubic HP-model within $\frac{3}{8}$ (0.38) of optimal.*

Theorem 3.9 (Newman 2002) *There is a linear-time approximation algorithm that folds an arbitrary HP-protein sequence in the 2D square HP-model within $\frac{1}{3}$ (0.33) of optimal.*

Theorem 3.10 (Newman 2002) *There exist HP-sequences S whose optimal fold in the 2D HP-model satisfies:*

$$OPT_{2D}(S) \leq (1 + o(1))\frac{\mathcal{C}_{2D}(S)}{2}$$

Therefore, any algorithm that uses the bounding argument of Theorem 3.7 to obtain a mathematically guaranteed approximation ratio cannot approximate better than $\frac{1}{2}$ of optimal.

Theorem 3.11 (Newman-Ruhl 2004) *There is a linear-time approximation algorithm that folds an arbitrary HP-protein sequence in the 3D square HP-model within* 0.37501 *of optimal (improving on $\frac{3}{8} = 0.3750$).*

Theorem 3.12 (Mauri-Pavesi-Piccolboni 1999) *There is a cubic-time approximation algorithm that folds an arbitrary HP-protein sequence in the 2D square lattice HP-model within $\frac{1}{4}$ (0.25) of optimal.*

3. 2D Square and 3D Cubic Lattice with Diagonals (so-called Extended Lattices)

   Theorem 3.13 (Bokenhauer-Bongartz 2007) *There is a linear-time approximation algorithm that folds an arbitrary HP-protein sequence in the extended 2D cubic lattice HP-model within $\frac{15}{26}$ of optimal. There is a linear-time approximation algorithm that folds an arbitrary HP-protein sequence in the extended 3D cubic lattice HP-model within $\frac{5}{8}$ of optimal.*

4. Face Centered Cubic Lattice (FCC)

3.2.2 Approximation Algorithms for Side-Chains Models

Theorem 3.14 (Hart-Istrail 1997) *There is a linear-time approximation algorithm that folds an arbitrary HP-protein sequence in the extended 3D FCC lattice HP-model within $\frac{31}{36}$ (0.86) of optimal.*

Theorem 3.15 (Heun 2003) *There is a linear-time approximation algorithm that folds an arbitrary HP-protein sequence in the extended cubic lattice side chain HP-model within $\frac{59}{70}$ (0.84) of optimal.*

5. Hexagonal Lattice

Theorem 3.16 (Jiang-Zhu 2005) *There is a linear-time approximation algorithm that folds an arbitrary protein sequence in the 2D hexagonal lattice HP-model within* $\frac{1}{6}$ *(0.17) of optimal.*

3.2.3 Approximation Algorithm for an Off-Lattice Model

Theorem 3.17 (Hart-Istrail 1997) *There is a linear-time approximation algorithm that folds an arbitrary HP-protein sequence in the Tangent Spheres side-chain HP-model within* $\frac{31}{36}$ (0.86) *of optimal.*

NAME	YEAR	LATTICE TYPE	MODEL TYPE	APPROX. RATIO
Jiang-Zhu	2005	2D Hexagonal Lattice	Side-Chains	$\frac{1}{6}$ (0.17)
Hart-Istrail	1995	2D Square Lattice	Linear Chains	$\frac{1}{4}$ (0.25)
Mauri-Pavesi *et al*	1999	2D Square Lattice	Linear Chains	$\frac{1}{4}$ (0.25)
Newman	2002	2D Square Lattice	Linear Chains	$\frac{1}{3}$ (0.33)
Newman-Ruhl	2004	3D Cubic Lattice	Linear Chains	$\frac{3}{8}$ (0.3750)
Hart-Istrail	1995	3D Cubic Lattice	Linear Chains	$\frac{3}{8}$ (0.38)
Batzoglou-Decatur	1996	2D Triangular Lattice	Linear Chains	$\frac{1}{2}$ (0.5)
Agarwala-Batzoglou *et al*	1997	2D Triangular Lattice	Linear Chains	$\frac{6}{11}$ (0.55)
Agarwala-Batzoglou *et al*	1997	3D Triangular Lattice	Linear Chains	$\frac{44}{75}$ (0.59)
Batzoglou-Decatur	1996	3D Triangular Lattice	Linear Chains	$\frac{9}{15}$ (0.60)
Bokenhauer-Bongartz	2007	3D Cubic w/ Diagonals	Linear Chains	$\frac{5}{8}$ (0.62)
Heun	2003	2D Square w/ Diagonals	Linear Chains	$\frac{59}{70}$ (0.84)
Hart-Istrail	1997	FCC Lattice	Side-Chains	$\frac{31}{36}$ (0.86)
Hart-Istrail	1997		Off-Lattice	$\frac{31}{36}$ (0.86)

3.3 Combinatorial Optimization Methods

There are a number of powerful combinatorial optimization methods in this area (see [79] for a survey). The most successful method is based on constraint programming and is due to Backofen and his collaborators [11, 12, 13, 17, 15]. Backofen and Will present a constraint-based method for finding optimal folds in the three-dimensional cubic and face-centered cubic lattices [14, 16]. Backofen has also been able to find upper bounds on the number of contacts on the FCC lattice in the HP-model [12]. A number of other linear programming based methods have been developed as well [26, 107, 28]. An interesting result about an exact exponential algorithm for protein folding is the following. Fu and Wang give a divide-and-conquer approach based on geometric separators to design a $2^{O(n^{1-\frac{1}{d}\log n})}$ exact algorithm on a d-dimensional grid [51].

4 Protein Folding: Computational Complexity

"The exactness of mathematics is well illustrated by proofs of impossibility. When asserting that doubling the cube ... is impossible, the statement does not merely refer to a temporary limitation of human ability to perform this feat. It goes far beyond this, for it proclaims that never, no matter what, will anybody ever be able to [double the cube]. No other science, or for that matter no other discipline of human endeavor, can even contemplate anything of such finality."
Mark Kac and Stan Ulam 1968

"For a quarter of a century now NP-completeness has been computer science's favorite paradigm, fad, punching bag, buzzword, alibi, and intellectual export... pervasive and contagious." Christos Papadimitriou 1995

An important question to address is the inherent complexity of computing the lowest-energy fold of a protein sequence in a given model. In this section, we survey computational complexity results for protein-folding lattice models. We present a series of theorems establishing that in a variety of models, both well studied biochemical models and ad-hoc models introduced for the purpose of the mathematical proof, finding the minimum-energy fold for a protein sequence is computationally intractable, i.e., NP-complete. These types of mathematical results were well received in the computational protein-folding community, as on the one hand they provided vindication for the apparent intrinsic difficulties and slow progress in developing accurate protein-folding algorithms, and on the other hand, they identified rigorous islands of well understood computational bottlenecks.

An *NP-hardness* theorem for a computational problem classifies its computational complexity as being as difficult to solve as certain problems in a much celebrated and well studied class of problems, icons of computational difficulty, referred to as computationally intractable; this class of problems includes the Traveling Salesman Problem, the Boolean Satisfiability of Propositional Logic, and the Set Cover Problem in graph theory [52]. The famous question whether NP-hard problems can be solved in polynomial time, the so-called P vs. NP Problem, is one of the seven magnificent Clay Mathematics Institute Millennium Million-Dollar Prize problems: the Birch and Swinnerton-Dyer Conjecture, the Hodge Conjecture, the Navier-Stokes Equations, the P vs. NP Problem, the Poincaré Conjecture, the Riemann Hypothesis, and the Yang-Mills Theory [3].

4.1 NP-completeness: from 10^{300} to 2 Amino Acid Types

The NP-hard theorems presented in this section focus on a set of computational problems in the protein-folding context. Most of them have to do with the packing of hydrophobic cores, but other computational aspects of protein-structure prediction are considered as well. The first NP-hard result, from 1992, is by Ngo and Marks [89], followed by Fraenkel [49, 50] and Unger and Moult [102]. The models used in these papers were introduced ad hoc for the purpose of showing that protein-folding-like computational tasks were similar in difficulty to well-studied NP-complete problems. In fact, the NP-completeness results that followed converged in model features toward parameters matching real proteins in biophysical models. The series culminated with the NP-completeness results for Dill's HP-model in 1998: Crescenzi, Goldman, Piccolboni, Papadimitriou and Yannakakis showed the 2D square-lattice model is NP-complete [35], while Berger and Leighton showed the 3D cubic lattice model is NP-complete [21]. This series of proofs used reductions from a variety of well-known NP-complete problems – SATISFIABILITY, PARTITION, OPTIMAL LINEAR ARRANGEMENT, 3D MATCHING, NOT-ALL-EQUAL 3SAT, P3SAT, MAX-CUT, MAX-3SAT, HAMILTONIAN PATH, BIN PACKING [52] – showing in the realm of mathematical proofs tremendous breadth and depth in how computational difficulty enters into the protein-folding modeling. The following gallery of theorems is presented without formal definition of the model involved or associated computational problem found to be NP-hard. The reader will find in the cited literature the details of the model, problem, and proof.

4.2 NP-completeness: Protein Folding in Ad-Hoc Models

The first proofs of NP-completeness were obtained for models of protein folding designed especially for the proof. In 1992 Ngo and Marks published the first NP-completeness results for 3D models for "commonly encountered energy-minimization tasks," modeling the geometry of backbone conformation of structure prediction for idealized carbon chains with

tetrahedral bond geometry. Although compactness of the chains is not the focus, the problems entail modeling interesting aspects of the folding process. A shortcoming of the results is that the encoding of the problem instance is *exponential* in the size of the protein sequence. The reduction used is from the PARTITION problem [89].

Theorem 4.1 (Ngo-Marks 1992) *The following three problems are NP-hard: DIAMOND LATTICE PATH (DLP), ENDPOINT CONSTRAINT POLYMER STRUCTURE PREDICTION (ECPSP), POLYMER STRUCTURE PREDICTION (PSPS).*

In 1993 and 1994, Fraenkel published NP-hard results for both 2D and 3D models capturing aspects of protein structure prediction. The protein model is not a chain but a graph and adjacency of amino acids in the chain and the self-avoiding walk requirement of the backbone are enforced by the optimization of the objective function. It is interesting to note modeling aspects related to potential functions involving all atom Coulomb-like energy minimization, Euclidian distance, and sum over all pairwise interactions of amino acids in the protein. The reductions is from 3D MATCHING [52] and ISING Model [49, 50].

Theorem 4.2 (Fraenkel 1993, 1994) *The 2D and 3D MINIMUM FREE ENERGY CONFORMATION OF PROTEIN (MEP) are NP-hard.*

In 1993, Unger and Moult established the NP-hardness for an all atom pairwise energy function depending on the types of amino acids in each pair and on the distance between the amino acids, on the 3D cubic lattice with diagonals on faces such that each lattice node has 26 neighbors. The self-avoiding walk restriction of the backbone was not part of the model but was enforced by penalties. The reduction used was from OPTIMAL LINEAR ARRANGEMENT [102].

Theorem 4.3 (Unger-Moult 1993) *The DPF PROTEIN FOLDING problem is NP-hard.*

Paterson and Przytycka constructed in 1996 a model with an unbounded number of amino acids (their number grows with the protein sequence length) that for the first time resembled the HP-model in that contacts were only between identical types of amino acids. They demonstrated NP-hardness of several problems including a "multi-string folding problem," a combinatorial jewel called "crossover folding." Their reductions were from SATISFIABILITY, NOT-ALL-EQUAL 3SAT, P3SAT, showing how protein folding in their model can be viewed as "computing" circuits on Boolean inputs. [92]

Theorem 4.4 (Paterson-Przytycka 1996) *The following problems are NP-hard: 2D square and 3D cubic STRING-FOLD, and 2D square MULTISTRING-FOLD.*

NP-completeness results are rarely robust. Adding 1 to the objective function can transform the problem from tractable to intractable and vice versa. With our incomplete knowledge about protein-folding energy function structure, it is interesting to investigate robustness of such NP-completeness results when one varies the parameters of the model. Hart and Istrail in 1997 gave computational intractability results for the protein-folding problem on lattices that are robust and can be applied to a variety of energy functions [61].

Theorem 4.5 (Hart-Istrail 1997) *There exist models for protein folding in which the NP-completeness of finding the lowest-energy conformation is invariant (1) when changing model lattice types among Bravais lattices, (2) when the energy function includes 3D distances between amino acids in the fold, and (3) when the model is either linear-chain or side-chain.*

In 1997 Nayak, Sinclair, and Zwick presented the first NP-hardness results for string folding in an HP-like model with a *finite* number of amino acid types. "Finite" is in fact 10^{300}, but this was the first result in which NP-completeness was proved for a finite alphabet of amino-acid types. The paper contains a variety of very interesting results involving spatial coding theory, a variant of the classical error-coding theory redesigned to model the geometry of folding strings. They also consider multi-string folding problems. They obtain hardness of approximation (even to approximate close to optimal is NP-complete) results for the string-folding problems. The reductions used were from MAX-CUT, MAX-3SAT [86].

Theorem 4.6 (Nayak-Sinclair-Zwick 1999) *The MAX-1 FOLD Problem is NP-hard.*

The first NP-completeness focused on side-chain packing with a rotamer library was published by Akutsu in 1999. Although for an artificial model, the method of proof deals for the first time with side-chain packing and perturbation modeling issues. The proof uses a reduction from 3SAT [6].

Theorem 4.7 (Akutsu 1999) *The PROTEIN SIDE-CHAIN PACKING WITH A ROTAMER LIBRARY is NP-hard.*

Atkins and Hart published in 1999 the first NP-completeness proof for an HP-like model with (humanly) finite size, namely 12 amino acid types. The contacts are only between identical types, and the lattice used is the 3D cubic lattice. The paper also contains a rigorous proof of the known basic (part of the folklore) lemma: The set of lowest-energy folds of the sequence H^{N^3} in the 3D HP-model consists of all the $N \times N \times N$-cubes. The proof contains a number of interesting constructions including the ability to program subcubes of independent energy in Lego-like cube-folding space. The reduction is from HAMILTONIAN PATH [10].

Theorem 4.8 (Atkins-Hart 1999) *The (A,M)-PROTEIN FOLDING ((A,M)-PF) problem is NP-hard.*

4.3 NP-completeness: Protein Folding in the HP-Model

The series of results described in the previous section culminated in 1998 with the definitive results about the NP-hardness of the HP-model both on the 2D square lattice and on the 3D cubic lattice. Both proofs are quite involved and provided a "phase transition" at the frontier between theory and practice: the finality of the mathematical proof of computational intractability of the HP-model, one of the most studied protein-folding models, introduced by Ken Dill in 1985 [37].

Constructing optimal folds for the 2D square HP-model was shown NP-complete by Crescenzi, Goldman, Piccolboni, Papadimitriou and Yannakakis in 1998. They consider both multi-string and single-string folding problems. The reduction from the HAMILTONIAN CYCLE that they construct uses the Trevisan codes that are powerful mappings of a graph on the hypercube so that the adjacency vs. non-adjacency of nodes in the graph is related to their Hamming distance on the hypercube [35].

Theorem 4.9 (Crescenzi-Goldman-Piccolboni-Papadimitriou-Yannakakis 1998) *The 2D STRING FOLDING PROBLEM in the HP-model is NP-hard.*

Berger and Leighton showed in 1998 that the protein-folding problem in the 3D cubic HP-model is NP-hard. Their proof involves a reduction from BIN PACKING, which is a

strongly NP-complete problem. They also provide a rigorous proof of the basic lemma: "The set of lowest-energy folds of the sequence H^{N^3} in the 3D HP-model are the $N \times N \times N$-cubes." Their proof uses the classical result in geometry that the number of grid points p in a configuration with A, B, C points aligned along the X, Y, Z axes, respectively, is at most $\sqrt{ABC}$.

Theorem 4.10 (Berger-Leighton 1998) *The following two problems are NP-complete:*

- *The PERFECT HP STRING FOLD problem, of finding, for an HP sequence containing n^3 Hs, a fold in the 3D cubic lattice for which the H's are perfectly packed into an $n \times n \times n$ cube, is NP-complete.*
- *The HP STRING-FOLD problem, of folding an HP-protein sequence in the 3D cubic HP-model, is NP-hard.*

5 Self-Avoiding Walks, Statistical Mechanics and Contact Map Overlap

We present a set of rigorous results for fold alignment of self-avoiding walks in 2D and 3D space (not restricted to lattices). Such alignments focus on contacts of pairs of amino acids properly defined and on the set of all such contacts, called the *contact map.* The similarity measure used for fold alignment is the *Contact Map Overlap.* It turns out that the ability to provide rigorous algorithms for fold alignment led to a decomposition theorem for self-avoiding walks in contact substructures, which happen to correspond on lattices to protein secondary structure building blocks. These in turn, have connections with the combinatorics theory developed by Mike Waterman and collaborators, which provide leads towards rigorous statistical mechanics results for HP-models.

We focus attention here first on contact maps of protein folds and on the contact-map overlap measure of fold similarity. This measure has been the basis of practical and rigorous algorithms for structure comparison and analysis [56, 80, 25, 24]. The contact map of a protein fold is a graph that captures the pattern of contacts in the fold. There are various concepts of contacts. In the contact map of a protein fold for the HP model, there is a vertex for each amino acid and an edge between a pair of amino acids i and j when they form a contact in the fold. Protein folds are represented as self-avoiding curves with the amino acids represented as points on the curves. A contact in a given fold is typically defined to be a pair of amino acids with distance smaller than a given threshold.

5.1 1 SAW = 2 STACKs + 1 QUEUE

We first show that the contact map of a self-avoiding walk can be decomposed in two "stacks" and one "queue." It turns out that stacks (reminiscent of α helices) correspond to chains in partial order structures and are slight generalizations, viewed as graphs, of the RNA secondary structures. From a dual point of view, queues (reminiscent of β sheets) correspond to anti-chains in partial order structures.

In the two-dimensional HP model, on lattice or off-lattice, the structure of the contact map of any protein fold can be characterized as follows.

Theorem 5.1 (Goldman-Istrail-Papdimitriou) *[56, 57] For any HP-sequence S, the set of contacts or the contact map of any two-dimensional fold of S (in fact of any self-*

avoiding walk with contacts between a set of points marked along the walk) can be decomposed into two stacks and one queue.

In contrast to the situation for the two-dimensional HP model, the contact map overlap problem in the three-dimensional HP model is not known to have a constant-factor approximation algorithm. In particular, it is possible to show that the same approach of decomposing the contact maps into stacks and queues cannot result in a constant-factor approximation guarantee.

Theorem 5.2 *There is a fold in the three-dimensional HP model such that any decomposition of the contact map into stacks and queues requires a collection of* $\Omega(\sqrt{n})$ *stacks and queues.*

It is interesting to note that in studies done in [4], protein structure contact maps from PDB were found to be decomposable into a combined number of stacks and queues of size about 15.

Decomposing PDB protein structures into stacks and queues. Algorithms for the decomposition of the contact maps of PDB protein structures into Stacks and Queues are of interest. We conjecture a combined number much lower than the worst-case $\Omega(\sqrt{n})$ theoretical result.

5.2 Schmitt-Waterman Contact Trees

Figure 10: Stack

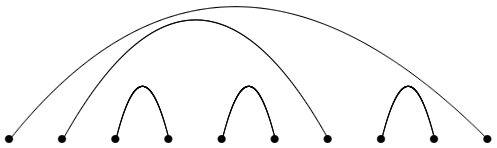

Figure 11: Schmitt-Waterman Contact Tree

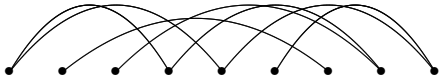

Figure 12: Queue

Waterman's work [96, 108] provided explicit analytical formulas for counting the total number of RNA structures of a given length, which could provide insights into computing rigorous approximations of the partition function of protein folding in HP models via the above decomposition.

If one considers the problem of finding an analytical closed form for the partition function of the HP-model in 2D, the above decomposition can be coupled with the following result of Schmidtt and Waterman [96] to lead to progress towards approximate counting of the contact substructures of folds. Related combinatorial theory was developed in [109, 108].

Theorem 5.3 (Schmitt-Watermann) *[96] There is a one-to-one bijection between RNA secondary structures and Schmitt-Waterman contact trees that can be completely enumerated in analytical closed form.*

Note that the Schmitt-Waterman contact trees are particular versions of stacks when the degree of each vertex is equal to 1. The problem of decomposing contact maps into building blocks has a parallel literature with similarity with the above results e.g., in statistical thermodynamics of double-stranded polymer molecules [30, 32], in protein core assembly [45] and in a knot theory approach to protein symmetries [31]. Note that the duality of the concepts of stack and queue in the partial order context mentioned could lead to dual enumerations techniques that could provide overall approximate counting results for partition functions of HP-models.

5.3 Fold Alignment by Contact Map Overlap

Protein-fold alignment is an important problem with applications to classifying known folds, predicting new folds, and judging the quality of prediction algorithms [44, 85, 68, 91]. In order to judge the quality of protein-folding algorithm, it is necessary to define a measure of fold similarity. Several such measures have been studied, including the root-mean-square-deviation (RMSD) measure [106, 17], the distance-map-similarity measure [69], and the contact-map-overlap measure [65, 55]. The *contact-map-overlap problem* is the problem of finding an order-preserving map of amino acids from one protein to the second protein that maximizes the number of common contacts. More formally, for protein sequences $S = \{s_1, s_2, \ldots s_n\}$ and $T = \{t_1, t_2, \ldots t_m\}$, the problem is to find an injective map $f : \{1, 2, ..., n\} \rightarrow \{1, 2, ..., m\}$ maximizing the sum

$$\sum_{i,j} e_S(i,j) e_T(f(i), f(j)).$$

In this sum, $e_S(i,j) = 1$ if (s_i, s_j) is a contact in sequence S and 0 otherwise; similarly for $e_T(k,l)$.

Although computing this measure exactly is NP-complete [56], it has been computed with great accuracy for PDB protein sequences of length about 100 using the integer linear programming techniques of Lagrangian relaxation and branch and cut [24, 25]. Some heuristics for computing this measure have also been shown to be effective, including local search and genetic algorithms. An important feature of this measure is its robustness with respect to the choice of threshold in defining amino acid contacts.

Since the contact-map comparison problem can be solved in polynomial time for stacks and queues through dynamic programming, this results in the following approximation algorithm.

Theorem 5.4 (Goldman-Istrail-Papdimitriou) *[56] There is an $O(n^6)$ time approximation algorithm within $\frac{1}{3}$ of optimal for the protein contact map problem in the two-dimensional HP model.*

More recently, the following new results show a tradeoff between time complexity of the algorithms and the approximation ratio.

Theorem 5.5 (Agarwal-Mustafa-Wang 2007) *[4] There is an $O(n^3 \log n)$ approximation algorithm for the maximum contact map overlap for two self-avoiding walks in 2D within $\frac{1}{6}$ of optimal.*

The same set of results include a theorem that shows the computational complexity of the problem in 3D.

Theorem 5.6 (Agarwal-Mustafa-Wang 2007) *[4] The problem of finding the maximum contact map overlap of two 3D self-avoiding walks in NP-hard.*

6 The Protein Side-Chain Self-Assembly Conjecture

In this section we propose a conjecture and give some mathematical evidence of its validity. We also show connections with the Kepler Sphere Packing Conjecture via a related problem for side-chain HP-models called the bi-spheres packing problem. The *Protein Side-Chain Self-Assembly Conjecture* asserts that as far as the optimal folding is concerned, the backbone is not essential. That is, the set of hydrophobic side-chains of a protein, if disconnected from the backbone, would "self-assemble" in a packing similar to their native packing achieved when they are connected to the backbone. We make this conjecture mathematically precise for lattice and off-lattice side-chain HP-models.

In the off-lattice Tangent Spheres side-chain HP-model [64], all amino acids are represented by identical spheres, and we use the following color coding scheme: backbone spheres are green, hydrophobic spheres are red and hydrophilic spheres are blue, see Figure 16. (The green spheres are drawn smaller to help see the hydrophobic core which has a biplane structure.) Hydrophobic contacts are made when two red spheres are tangent. One can consider a side-chain as a *bi-sphere,* that is a green sphere connected by an edge (alternatively the two identical sized spheres could just be tangent) to a red sphere – the hydrophobic bi-sphere, or a green sphere connected to a blue sphere – the hydrophilic bi-sphere. A protein has its set of hydrophobic bi-spheres and its set of hydrophilic bi-spheres. Given a set of n hydrophobic bi-spheres a contact is made when two red spheres (of two different bi-spheres) are tangent. Consider now the optimal arrangement of the set of n hydrophobic bi-spheres, which we will call the "self-assembly" configuration; that is, the arrangement that maximizes the number of contacts. We denote the optimal number of contacts by $BSA(n)$. Similar considerations can me made for lattice side-chain HP-models. We call the side-chains now "bipoles" to distinguish the lattice terminology from the off-lattice one. A *bipole* is just an edge with the two end-point nodes labeled green-red or green-blue.

Theorem 6.1 (SELF-ASSEMBLY BOUND INEQUALITY) *For a protein sequence S with n hydrophobic side-chains in an off-lattice (lattice) side-chain HP-model, let $OPT(S)$ be the number of contacts in an optimal fold of S. For an off-lattice (lattice) HP-model, let $BSA(n)$ be the number of contacts in the optimal hydrophobic bi-sphere (bipole) self-assembly of n bi-spheres (bipoles).*

Then the following inequality holds:

$$BSA(n) \geq OPT(S)$$

Therefore, if a fold F of a protein sequence S, with n hydrophobic side-chains, has $BSA(n)$ contacts, then F is an optimal fold for S.

The Protein Side-Chain Self-Assembly Conjecture.

- *For every protein sequence from PDB, the number of contacts of the optimal fold in the side-chain HP-model on the FCC lattice equals the number of contacts in the optimal self-assembly structure of the set of hydrophobic bipoles of the protein. Moreover, the hydrophobic core of its optimal fold is identical to the hydrophobic bipoles self-assembly structure.*

- *For every protein sequence from PDB, the number of contacts of the optimal fold in the (3D) Tangent Spheres side-chain HP-model equals the number of contacts in the optimal self-assembly structure of the set of hydrophobic bi-spheres of the protein. Moreover, the hydrophobic core of its optimal fold is identical (up to small perturbations preserving the tangent spheres structure) to the hydrophobic bi-spheres self-assembly structure.*

6.1 The Kepler Conjecture and Bi-Sphere Packing

The problem of sphere packing in 3D space can be stated as follows. Given an infinite set of spheres with identical radii, find an arrangement for the entire 3D space that has the highest density (minimizes the amount of unused empty space among them). For this type of sphere packing with highest density for the entire infinite 3D space, Kepler conjectured in 1611 [77] that the densest arrangement is achieved when the spheres are placed as vertices in a Face Centered Cubic (FCC) lattice; Gauss proved in 1840 [53] that indeed among lattices the FCC lattice provides the densest packing. The off-lattice problem, i.e., with no restriction to a regular lattice arrangement, remained open almost till the end of the last century. In 1997, Thomas Hale [59] provided the proof of the Kepler Conjecture that is considered today by experts to be "99% certain." His proof, parts of which involved computer-checked subproblems, shows that FCC is indeed the densest packing even if non-lattice arrangements are considered as well. The same "best off-lattice is still the lattice" situation was found for the densest packing in 2D space as well for the circle-packing problem, where one wants the highest density of circles with identical radii. Lagrange proved in 1773 that the hexagonal packing is the best among the all lattices, and in 1953 Fejer-Toth showed that the densest off-lattice solution is still the hexagonal lattice packing. For finite space regions in 3D, however, FCC is no longer always the densest packing. The same is true for 2D, where for some finite regions, non-hexagonal packing can be denser. For the major reference text on sphere packing see Conway and Sloabe [34].

Densest Packing vs. Maximum Number of Contacts. Two measures are of interest regarding packing of spheres, and we consider next the problem of bi-sphere packing. The first measure is *packing density* as defined above in the Kepler Conjecture: minimizing the unused space in the packing. The second is maximizing the number of contacts, where a *contact* for a pair of spheres is defined here as in the kissing-number problem, i.e., the two spheres are tangent to each other. A natural question is whether or not the optimal packing arrangements for these two measures are equivalent. The answer is "no" for finite regions of 2D and 3D space and "yes" for infinite 2D and 3D space.

6.1.1 The Bi-Sphere Packing Problem

The problem we consider now is that of packing in space pairs of tangent spheres bound together and colored; one sphere of each pair is colored red and the other green. We call such an object a *bi-sphere* in off-lattice space and *bipole* or "side-chain edge" in the lattice model where one vertex of the bipole is the backbone vertex where it is hooked, and the other vertex is the amino-acid type; red corresponds to the hydrophobic amino acid, green represents the backbone, and blue represents the hydrophilic amino acid type. A *bi-sphere contact* is formed when two bi-spheres are arranged so that their red spheres touch. We are interested here in bi-sphere configurations that achieve the maximum number of bi-sphere contacts. We call such a optimal configuration an *self-assembly.* The self-assembly problem was studied in [72], where the *biplane problem* was introduced.

The analogue to the Kepler problem for bi-spheres is the *Bi-sphere Packing Problem*, formulated as follows. Given a set of bi-spheres, find the 3D space arrangement, or self-assembly, achieving the maximum number of bi-sphere contacts (red-red sphere contacts). The *FCC Biplane Packing Configuration* proposed by Istrail et al. 2000 [72] is the following: the optimal self-assembly of the bi-spheres is placed on four consecutive hexagonal planes (numbered 1, 2, 3, and 4) of the FCC lattice; the red spheres are arranged on planes 2 and 3, on two (almost) equal-shaped regions facing each other; the green-sphere mates of the red spheres in plane 2 red are placed on plane 1, while the green-sphere mates of the red spheres in plane 3 are placed on plane 4. See Figure 16 where the corresponding hydrophobic bipoles of a protein are arranged in the conjectured biplane configuration. The *3D Cubic Lattice Biplane Configuration* for a set of n bipoles on the 3D cubic lattice is obtained by placing the red spheres in two rectangles of side length equal or differing by one onto two consecutive parallel planes, with the red spheres mapped to adjacent planes of the 3D cubic lattice and the green mate spheres on the adjacent unoccupied planes. The following are conjectures on the optimality of the biplane configurations.

3D FCC Lattice Biplane Packing Conjecture *The optimal bi-sphere packing off-lattice or bipole packing in the FCC lattice is achieved by the FCC biplane configuration.*

3D Cubic Lattice Biplane Packing Conjecture *The optimal bipole packing in the 3D cubic lattice is achieved by the 3D cubic biplane configuration.*

For a lattice L, the *HP-self-assembly L model* is defined as follows. Given a set of n bipoles (red-green edges), we are interested in the self-assembly configuration of the set of bipoles. Each bipole is embedded as a disjoint edge of the lattice L and the self-assembly is an arrangement that has the maximum number of red-red contacts (two red vertices from two bipoles connected by a lattice edge). Istrail et al. show lower bounds for the bipole self-assembly problem, giving rise to the first result that biplane configurations are within a small percent of the optimal energy configuration [72].

Figure 13(a) shows an example of a packing of 13 bipoles resulting in 11 contacts. However, a packing of 13 bipoles with 20 contacts is shown in Figure 13(b) (the figure denotes the (red) side-chain amino acids by filled circles and backbone molecules (green) by unfilled circles).

(a) Arrangement of 13 bipoles

(b) Optimal packing of 13 bipoles

Figure 13: Bipole packing configurations on the 2D square lattice

6.2 Optimal Bipole Packing on 2D Square Lattice

In the following theorem, we show that the pattern in Figure 13(b) can be generalized to obtain the exact optimal configuration for bipole packing, the self-assembly of the n bipoles, in the two-dimensional square lattice.

Theorem 6.2 (Optimal Bipole Packing on the 2D Square Lattice) *For a set of n bipoles, the maximum number of hydrophobic contacts in a packing in the two-dimensional square lattice is $\lfloor \frac{3n-4}{2} \rfloor$. This bound is tight and is achieved for the configuration of hydrophobic molecules arranged in two lines, as shown in Figure 13(b).*

Proof. For any packing P of the n bipoles, consider drawing a vertical and horizontal line (parallel to the x and y-axes) through each backbone molecule. Then each backbone molecule p divides the lattice into four quadrants $Q_1(p), Q_2(p), Q_3(p)$, and $Q_4(p)$ as shown in Figure 14.

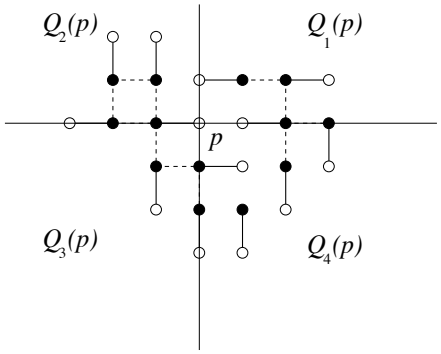

Figure 14: Quadrants $Q_1(p), Q_2(p), Q_3(p)$, and $Q_4(p)$

Note that we include the boundary horizontal and vertical lines in their respective quadrants. For each point p, let $q_i(p)$ denote the number of backbone molecules other than p in $Q_i(p)$. Note that a backbone molecule q may belong to two quadrants of p if it lies on a boundary line between the two quadrants. We first show that for each $1 \leq i \leq 4$, there exists a point p_i with the property that $Q_i(p_i)$ does not contain any backbone or side-chain molecules other than those of p_i, i.e. $q_i(p_i) = 0$.

For a fixed i, $1 \leq i \leq 4$, suppose there is no point p_i with $q_i(p_i) = 0$. Consider a point p with the smallest value $q_i(p)$ (by assumption $q_i(p) > 0$). Then there exist points in $Q_i(p)$ and any such point p' satisfies $q_i(p) > q_i(p')$. This contradicts the choice of p, and therefore $Q_i(p)$ must be the empty set.

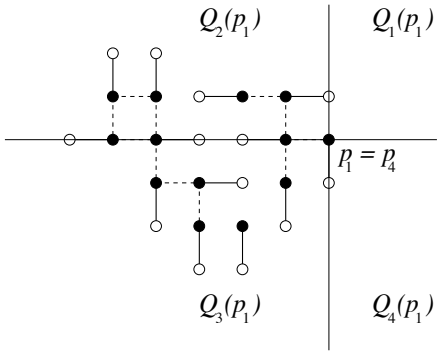

Figure 15: Quadrants $Q_1(p), Q_2(p), Q_3(p)$, and $Q_4(p)$

It may be possible for $p_i = p_j$ for $1 \leq i < j \leq 4$ (such as in Figure 15), but there must be at least two distinct values among p_1, p_2, p_3, p_4. Each backbone vertex p has three adjacent

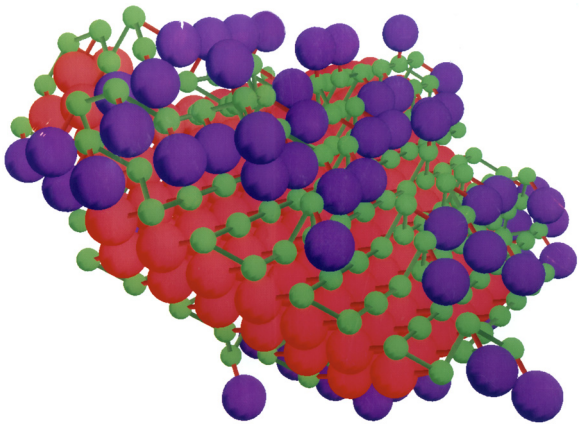

Figure 16: A PDB protein represented in the HP-side chain model, folded near optimally (98%) on the FCC lattice. Red is hydrophobic, Blue is hydrophilic, Green is backbone.

positions for potential contacts, the fourth position occupied by the side-chain molecule. If not all three positions are occupied by backbone molecules, there is a lost contact. Each empty quadrant $Q_i(p)$ leads to one lost contact with p. Therefore, there is a total of four lost contacts for any configuration, one for each of p_1, p_2, p_3 and p_4. The maximum number of contacts for any backbone molecule is 3, and taking into consideration the four lost contacts and double counting of each contact, the maximum number of contacts in a packing of n bipoles is therefore

$$\text{Maximum number of hydrophobic contacts } \leq \left\lfloor \frac{3n-4}{2} \right\rfloor.$$

Note that the arrangement of bipoles into two lines as shown in Figure 13(b) achieves exactly four lost contacts for n even and five lost contacts for n odd. This matches the upper bound above, proving the theorem. □

Optimal Bipole Packing and Optimal Bi-Sphere Packing Problems.

- As far as we know, the above proof of the optimality of bipole packing (self-assembly) on the 2D square lattice is the only one such proof available. It would be very interesting to obtain similar proofs for 2D hexagonal lattice, 3D cubic lattice and the FCC lattice. Similarly, for off-lattice, it would be important to find the self-assembly structures and prove their optimality for bi-sphere packing in 2D and 3D space.
- It is a low hanging fruit for combinatorial optimization methods to provide optimal solutions for bipole packing and bi-sphere packing for small sizes (25) of bipoles and bi-spheres. That would already be informative towards a number of conjectures in this paper. Computing optimal self-assembly configurations for minimum to large sizes (50-150) for the above set of lattices and for 2D and 3D space would definitely make an important contribution to this research area.

The bipole framework also gives rise to the following related problems, each addressing a different objective in the problem of packing bipoles.

Bipole Configuration Labeling Problem Given a fixed bipole packing of unlabeled bipoles, find an assignment for the bipole labels such that

(1) each bipole has one endpoint labeled H (representing the hydrophobic side chain) and one endpoint labeled B (representing the backbone molecule)

(2) the number of contacts between endpoints labeled H is maximized over all possible assignments

Bipole Configuration Local Moves Problem Given two bipole packings (with labeled hydrophobics and backbone vertices), define a set of local moves transforming one bipole packing to the other. Under the set of local moves defined, find the *minimum* number of such moves taking one bipole packing to the other.

7 Concluding Remarks

7.1 Discussion

Approximation Algorithms. It is remarkable that all the approximation algorithms presented in this survey have a very similar form to the first such approximation algorithm [60]; all algorithms use similar balancing points that have the power to create a significant number of contacts non-locally, where significant is defined as a fraction of the number of optimal contacts. This similarity in algorithmic design transcends issues such as lattice parity constraints in forming contacts in some of the most studied lattices. This balancing-point strategy resembles the Zipper and Assembly paradigm [41]. The near optimal time of almost all the algorithms (linear in the number of amino acids of the protein) of constructing folds close to the optimal resembles the hydrophobic collapse and molten globule stages of folding.

All the approximation algorithms presented apply for the entire class of HP-sequences; that is, they fold each and every binary sequence of H's and P's achieving in linear time the claimed guarantee. There is a fundamental limitation in such results, namely, only theorems that are true for the entire set of binary sequences could be proved. One natural line of research is to limit the class of sequences to a subclass that is defined to include large subsets of PDB protein sequences. For a restricted set of proteins sequences, better performance algorithms could be obtained. In fact, one such result is due to Huen [67] who obtains, for a subclass of protein sequences subject to some "natural" protein-like pattern restriction, an approximation ratio of 88.1%, improving over the best general approximation ration of 86%.

Computational Complexity. The first proofs of NP-completeness from 1992 and 1993, although for protein folding in ad hoc models, led to the some speculations that the problem might be "confronting science's logical limits" [27] or that biology "solves" NP-complete problems [48, 102]. As this surveys shows, NP-completeness is universal for the HP-models and their generalizations, and other models as well. In one case, the 2D HP-model on hexagonal lattices, the authors of an approximation algorithm for the model conjectured that finding the optimal fold is NP-hard, which would imply the model is more realistic for modeling folding [75].

2D Hexagonal Lattice HP-Folding Problem. A very interesting open problem is whether on the 2D hexagonal lattice HP-folding is NP-complete. It would be intriguing if a

polynomial-time algorithm were to exist for this model.

Statistical Mechanics. The results of Theorem 5.1 and as those of Agarwal et al. in Theorem 5.5 provide an analytical point of connection between self-avoiding contact maps and RNA contact maps. The results of Waterman [108], Waterman and Smith [109] and Schmitt and Waterman [96] provide about "half" (stacks) of the analytical characterization needed; if the second part (queues) – a natural dual in the partial order set structure of "contact inclusion" – could be analytically obtained together, it could lead to a rigorous approximation algorithm for computing the partition function of self-avoiding walks in 2D. The stacks (in their degree-one restriction) correspond to Schmitt-Waterman contact trees of RNA structures (and to α-helices on lattices) counted exactly in [109], while queues correspond to β-sheets on lattices.

Counting Contact Maps of Stacks and Queues Problems. There is the duality between stacks as chains in the partial order of "contact inclusion," while the queues correspond to anti-chains in the same partial order. This duality, via Dilworth Theorem, could lead to generalizations of the Schmitt-Waterman counting formulas for queues that would then provide approximation for counting results for partition functions via the decomposition theorem of [56, 57]. Are there generalizations of the Schmitt-Waterman counting formulas to stacks (in full generality) and to queues?

7.2 Five Problems With Solutions Within Reach

The following problems represent research directions related to the results which we believe are within reach of obtaining provable results extending those in this survey.

1. **PROBLEM 1: Sequence discrimination: Not all sequences should be treated equally.** The present survey indicates that the best approximation algorithms ratios presented are quite hard to improve due to the relative weakness of the upper bound estimates of the number of native contacts, or the general results that can be obtained when one considers HP-models where all the binary H-P sequences are valid protein sequences. Obviously, this presents a challenge of considering definitions of patterns of natural proteins that would include large subsets of the PDB protein sequences. One such approach that shows promise is [67]. Work on sequence patterns related to folding pathways requirements inspired by studies of protein misfolding and aggregation showed for example that "frequencies of amino-acid strings in globular proteins sequences indicate suppression of blocks of consecutive hydrophobic residues" [73, 97, 47, 93, 2].

2. **PROBLEM 2: Equal rights for lattices: All lattices should be treated equally.** The master approximation algorithm 3.1 [62] shows that achieving an approximation algorithm within a constant of optimal is a universal property of crystallographic lattices. The proof involves showing that there is a general sublattice structure, responsible of the phenomenon, present in every 3D lattice analogous to the concept of "completeness" used in the "NP-complete" concept in computer science. It would be very interesting to obtain a similar result for the NP-completness of folding in HP-models across lattices. Such a result exists for example for the Ising model showing NP-completeness for each and every 3D lattice (in fact any non planar and translational invariant lattice) [71]. In particular, robust NP-completeness results incorporating general forms of energy functions would be especially effective in understanding the islands of tractability in the sea of computational intractability.

3. **PROBLEM 3: Fairness of potential-energy function versus protein structure voting rights: Individual amino acids interaction preference-values in protein structures versus the energy function social choice.** The thermodynamic hypothesis of Anfinsen [7] could be interpreted via voting theory [9] as a postulate on the existence of a "social choice," namely the postulated potential-energy function, having a universal form that analytically aggregates the set of all "interacting units." Such units are described by the information contained in protein structure data bases, such as the Protein Data Bank (PDB), about amino acid interactions, pairwise or more complex, for all such pairs or multi-way basic units of interaction.

   > *"[T]he true elegance of this consequence of natural selection was dramatized by the ribonuclease work since the refolding of this molecule after full denaturation by reductive cleavage of its four disulfide bonds ... required that only 1 of 105 possible pairings of eight sulfhyryl groups to form four disulfide linkages take place. ... to establish ... the "thermodynamic hypothesis." This hypothesis states that the three-dimensional structure of a native protein in its normal physiological milieu (solvent,* p*H, ionic strength, presence of other components such as metal ions or prosthetic groups, temperature, and other) is the one in which the Gibbs free energy of the whole system is lowest; that is, that the native conformation is determined by its totality of interatomic interactions and hence by the amino acid sequence, in a given environment."* Christian Anfinsen 1973 [7]

   Such interacting units represent the "individual values" of protein structures, and the postulated social choice potential-energy function should be such that it is consistent with the extraction of interaction units from protein structures in a "fair" way. Protein structures are the "individual voters" and the postulated potential-energy function is the social choice (although, actually, an all pairwise sum – if we assume pairwise potentials – of the individual pairwise interactions) extracted fairly from individual preferences. The major unresolved problem of finding statistical potentials for realistic folding applications, such as simulations and predictions, is at the heart of difficulty of the protein folding problem [74, 101]. The analogy with the Arrow Paradox or the Arrow Impossibility Theorem [9] could provide avenues for proving analogous mathematical impossibility theorems about potential-energy function inference scenarios. Such attempts are close to being impossibility proofs, without the extra difficulty to formalize assumptions used in the proof, e.g., the impossibility of pairwise potentials for protein folding models, e.g., [107]. The analogy with the voting theory framework could inspire also, more importantly, positive results, i.e., powerful algorithms for potential-energy inference. The work of von Neumann-Morgernstern [105] provided a concept of choice or preference of a statistical nature (this in contrast here with the deterministic concept of interaction) through a set of rules of "Axioms of Preference" that completely axiomatized the expected utility theory in economics. There are clear parallels between utility functions and potential-energy functions that could inspire a much needed phase transition towards novel rigorous methods of inference of such realistic folding potentials.

4. **PROBLEM 4: A combinatorial theory for balance (turning) points in protein sequences**. The approximation algorithms [60, 88] are both based on solutions of a set of combinatorial problems for finding balance points such that folds can be created that will assure a significant fraction of the number of native contacts. Follow up algorithms use similar ideas to iteratively apply these decompositions to improve

the packing and obtain folds closer to the estimated optimum. There is clearly room for developing a general combinatorial theory especially in concert with the similar biophysical counterpart, the Zipper and Assembly hypothesis for which there is also considerable empirical evidence on natural proteins. Again, patterns of natural protein sequences could lead to a more detailed combinatorics of balancing points as a way to build compact hydrophobic cores.

5. **PROBLEM 5: Freedom to self-assembly act.** The Protein Side-Chain Self-Assembly Conjecture, if true, could lead to the first HP-protein folding model for which a polynomial time approximation algorithm could be designed with better than 98% of optimal approximation ratio, see Figure 16 for an example. The model would be a side-chain HP-model on the FCC lattice for which a very close to optimal fold could be constructed in polynomial time, we guess $O(n^3)$, where n is the size of the protein sequence. The same approximation ratio performance would also be for off-lattice, i.e., for the Tangent Sphere Off-lattice HP-model [64]. The class of protein sequences would be restricted to class of sequences that will include a large set of protein sequences from the PDB. We presented in section 6 an overview of the ideas that indicate some validity and mathematical theorems that could lead to the desired "almost-optimal" algorithm.

8 The ProFolding Project

The following website contains benchmarks and algorithms for the HP-model.

`http://www.cs.brown.edu/~sorin/lab/pages/protein-folding.html`

9 Acknowledgments

The first author would like to express his deep gratitude to Ken Dill, whose work on protein folding introduced him to the area, and who over the years has been a strong source of support and inspiration. Thanks go to Ron Unger who challenged the first author in 1994 with the beautiful HP-model problem, which in turn, led to his work on protein folding. Many thanks for inspiring discussions and delightful collaborations go to Bill Hart, Jonathan King, Giuseppe Lancia, Bob Carr, Brian Walenz, John Conway, Russell Schwartz, Serafim Batzoglou, Ross Lippert, Bruce Hendrickson, Bill Camp, Al Hurd, and Fred Howes. Thanks to Mike Waterman and his colleague for organizing the Perseus Protein Folding Workshop at University of Southern California – the best workshop ever attended by the first author. Fumei Lam's work was supported by postdoctoral funding from the Brown University Office of the Vice President of Research. It is a great pleasure to acknowledge the editorial work of Erin Klopfenstein and Katrina Avery on this paper.

References

[1] Folding proteins fast (editorial on the first approximation algorithm for protein folding). *Science*, 269:1821, 1995.

[2] Protein adultery revealed! – mechanism of protein folding captured in computer simulation. *Science Daily*, February 25, 1999.

[3] Millenium prize problems. *Clay Mathematics Institute*, 2000.

[4] P. K. Agarwal, N. H. Mustafa, and Y. Wang. Fast molecular shape matching using contact maps. *J. Computational Biology*, 14:131–143, 2007.

[5] R. Agarwala, S. Batzogloa, V. Dancik, 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. *Journal of Computational Biology*, 4:276–296, 1997.

[6] T. Akutsu and S. Miyano. On the approximation of protein threading. *RECOMB 97*, pages 3–8, 1999.

[7] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[8] Aristotle. in nichomachean ethics. *Oxford University Press*, B15-20:1112, 1959.

[9] K. Arrow. Social choice and individual values. *John Wiley and Sons*, 1951.

[10] J. Atkins and W. E. Hart. On the intractability of protein folding with a finite alphabet of amino acids. *Algorithmica*, 25:279–294, 1994.

[11] 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 (PSB98)*, 2:387–398, 1998.

[12] R. Backofen. An upper bound for number of contacts in the HP-model on the facecentered-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, 2000.

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

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

[15] R. Backofen and S. Will. A constraint-based approach to fast and exact structure prediction in three-dimensional protein models. *Constraints*, 11:5–30, 2006.

[16] R. Backofen and S. Will. A constraint-based approach to fast and exact structure prediction in three-dimensional protein models. *Constraints*, 11(1):5–30, 2006.

[17] R. Backofen, S. Will, and P. Clote. Algorithmic approach to quantifying the hydrophobic force contribution in protein folding. *Altman R (ed), Pacific Symposium on Biocomputing, Honolulu*, pages 93–106, 2000.

[18] Z. Bagci, R. L. Jernigan R.L., and I. Bahar. Residue packing in proteins: uniform distribution on a coarse-grained scale. *Journal of Chemical Physics*, 116:2269–2276, 2002.

[19] Z. Bagci, R. L. Jernigan R.L., and I. Bahar. Residue coordination in proteins conforms to the closest packing of spheres. *Polymers*, 43(2):451–459, January 2002.

[20] C Bender. Bestimmung der grssten anzahl gleich kugeln, welche sich auf eine kugel von demselben radius, wie die brigen, auflegen lassen. *Archiv Math. Physik (Grunert)*, 56:302–306, 1874.

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

[22] J. U. Bowie, R. Luthy, and D. Eisenberg. A method to identify protein sequences that fold into a known 3-dimensional structure. *Science*, 253:164–170, 1991.

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

[24] A. Caprara, R. Carr, S. Istrail, G. Lancia, and B. Walenz. 1001 optimal PDB structure alignments: Integer programming methods for finding the maximum contact map overlap. *Journal of Computational Biology*, 11:27–52, 2004.

[25] A. Caprara and G. Lancia. Structural alignment of large-size proteins via Lagrangian relaxation. *RECOMB*, pages 100–108, 2002.

[26] R. Carr, W. E. Hart, and A. Newman. Bounding a protein's free energy in lattice models via linear programming. *Poster at RECOMB04*, 2004.

[27] J. L. Casti. Confronting science's logical limits. *Scientific American*, pages 102–105, 1996.

[28] M. Cebrian, I. Dotu, P Van Hentenryck, and P. Clote. Protein structure prediction on the face centered cubic lattice by local search. *Proceedings 23rd AAAI Conference on Artificial Intelligence*, pages 241–246, 2008.

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

[30] S. Chen and K. A. Dill. Statistical theormodynamics of double-stranded polymer molecules. *J. Chem. Phys.*, 103:5802–5813, 1995.

[31] S. Chen and K. A. Dill. Symmetries in proteins: a knot theory approach. *J. Chem. Phys.*, 104:5964–5973, 1996.

[32] S. Chen and K. A. Dill. Theory of conformational changes of double-stranded chain molecules. *J. Chem. Phys.*, 109:4602–4616, 1998.

[33] B. Cipra. Packing challenge mastered at last. *Science*, 281, 1998.

[34] J. H. Conway and N. J. A. Sloane. Sphere packings, lattices and groups, 3rd edition. *Springer-Verlag*, 1999.

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

[36] G. M. Crippen. Easily searched protein folding potentials. *Journal of Molecular Biology*, 260:467–175, 1996.

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

[38] K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29:7133–7155, 1990.

[39] 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.

[40] K. A. Dill, S. Banu Ozkan, T. R. Weikl, J. D. Chodera, and V. A. Voelz. The protein folding problem: when will it be solved? *Current Opinion in Structural Biology*, 17:2342–346, 2007.

[41] K.A. Dill, S. Banu Ozkan, M. Scott Shell, and T. R. Weikl. The protein folding problem. *Annual Review of Biophysics*, 37:289–316, 2008.

[42] Y. Duan and P. A. Kollman. Computational protein folding: From lattice to all-atom. *IBM Systems Journal*, 40:297–309, 2001.

[43] F. Dyson. Birds and frogs. *Notics of the American Mathematical Society*, 56:212–223, 2008.

[44] I. Eidhammer, I Jonassen, and William R. Taylor. Structure comparison and structure patterns. *Journal of Computational Biology*, 7:685–716, 2000.

[45] K. M. Fiebig and K. A. Dill. Protein core assembly processes. *J. Chem. Phys.*, 98:3475–3487, 1993.

[46] A.V. Finkelstein and B. A. Reva. Search for the stable state of a short chain in a molecular field. *Protein Engineering*, 5:617–624, 1992.

[47] S. A. In Focus. To fold or not to fold. *Scientific American.com*, 1999.

[48] A. S. Fraenkel. Deexponentializing complex computational mathematical problems using physical and biological systems. *Weizmann Inst. of Science*, TR CS90, 1990.

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

[50] A. S. Fraenkel. Protein folding, sping glass and computational complexity. *Third Annual DIMACS Workshop on DNA Based Computers*, pages 1–19, 1997.

[51] B. Fu and W. Wang. A $2o(n^{1-1/dlogn})$ time algorithm for d-dimensional protein folding in the hp-model. *Lecture notes in computer science, International Colloquium on Automata, Languages and Programming*, 2004.

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

[53] C. F. Gauss. Recension der 'untersuchungen uber die eigenschaften der positiven ternaren quadratischen formen' von ludwig august seeber. *Journal fur die reine und angewandte Mathematik*, pages 312–320, 1840.

[54] N. Go. Theoretical studies of protein folding. *Annual Review of Biophysics and Bioengineering*, 12:183–210, 1983.

[55] A. Godzik and J. Skolnick. Flexible algorithm for direct multiple alignment of protein structures and sequences. *Computer Applications in the Biosciences*, 10:587–596, 1994.

[56] D. Goldman, S. Istrail, and C. Papadimitriou. Algorithmic aspects of protein structure similarity,. *Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS99), IEEE Computer Society Press*, October 1999.

[57] D. G. Goldman. Algorithm aspects of protein folding and protein structure similarity. *Ph.D. Thesis, U.C. Berkeley*, 2000.

[58] L. Greengard. The rapid evaluation of potential fields in particle systems. *ACM Distinguished Dissertation*, MIT Press, Cambridge, MA, 1987.

[59] T. C. Hales. Sphere packings i and ii. *Discrete and Computational Geometry*, 17:1–149, 1997.

[60] W. E. Hart and S. Istrail. Fast protein folding in the Hydrophobic-Hydrophilic model within three-eighths of optimal (extended abstract). *Proceedings of 27th Annual ACM Symposium on Theory of Computation (STOC95)*, pages 157–168, 1995.

[61] W. E. Hart and S. Istrail. Robust proofs of NP-hardness for protein folding: General lattices and energy potentials. *Journal of Computational Biology*, 4:1–22, 1997.

[62] W. E. Hart and S. Istrail. Invariant patterns in crystal lattices: Implications for protein folding algorithms. *Proceedings of the 7th Conference on Combinatorial Pattern Matching (CPM96), Springer Lecture Notes in Computer Science*, pages 288–303, June 1996.

[63] W. E. Hart and A. Newman. The computational complexity of protein structure prediction in simple lattice models, in handbook of computational molecular biology. *CRC Press*, pages 1–24, 2001.

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

[65] T. F. Havel, G.M. Crippen, and I.D. Kuntz. Effect of distance constraints on macromolecular conformation II. simulation of experimental results and theoretical predictions. *Biopolymers*, 18:73–81, 1979.

[66] M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. J. Sippl. Identification of native protein folds amongst a large number of incorrect models: The calculation of low endergy conformations from potentials of mean force. *Journal of Molecular Biology*, 216:167–180, 1990.

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

[68] L. Holm, S. Kaariainen, P. Rosenstrom, and A. Schenkel. Searching protein structure databases with DaliLite v.3. *Bioinformatics*, 24:2780–2781, 2008.

[69] L. Holm and C. Sander. Mapping the protein universe. *Science*, pages 595 – 602, 2 August 1996.

[70] R. Hoppe. Bemerkung der redaction. *Archiv Math. Physik. (Grunert)*, 56:307–312, 1874.

[71] S. Istrail. Statistical mechanics, three-dimensionality and np-completeness: I. universality of intractability of the partition functions of the ising model across non-planar lattices. *Proceedings of the 32nd ACM Symposium on the Theory of Computing (STOC00), ACM Press*, pages p. 87–96, 2000.

[72] S. Istrail, A. Hurd, R. Lippert, B. Walenz, S. Batzoglou, J. H. Conway, and F. W. Peyerl. Prediction of self-assembly of energetic tiles and dominoes: Experiments, mathematics, and software. *Sandia National Labs Technical Report*, April 2000.

[73] S. Istrail, R. Schwartz, and J. King. Lattice simulations of aggregation funnels of protein folding. *Journal of Computational Biology*, 6, 1999.

[74] R. L. Jernigan and I. Bahar. Structure-derived potentials and protein simulations. *Current Opinion in Structural Biology*, 6:195–209, 1996.

[75] M. Jiang and B. Zhu. Protein folding in the hexagonal lattice in the hp model. *J. of Bioinformatics and Computational Biology*, 3:19–34, 2005.

[76] T. Jiang, Q. Cui, G. Shi, and S. Ma. Protein folding simulations of the hydrophobic-hydrophilic model by combining tabu search with genetic algorithms. *The Journal of chemical physics*, 119(8):4592–4596, 2003.

[77] J. Kepler. The six-cornered snowflake. 1611.

[78] J. King, C. Haase-Pettingell, and D. Gossard. Protein folding and misfolding. *American Scientist*, 90:445–453, 2002.

[79] G. Lancia. Mathematical programming in computational biology: an annotated bibliography. *Algorithms*, 1:100–129, 2008.

[80] G. Lancia, R. Carr, B. Walenz, and S. Istrail. 101 optimal PDB structure alignments: A branch-and-cut algorithm for the maximum contact map overlap problem. *Proceeding of the Fifth International Conference on Research on Computational Biology, (RECOMB 2001)*, pages 201–211, 2001.

[81] 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.

[82] M. Levitt and A. Wharshel. Computer simulation of protein folding. *Nature*, 253:694–698, 1975.

[83] A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga. A united-residue force field for off-lattice protein-structure simulations: I. functional forms and parmeters of long-range side-chain interaction potentials from protein crystal data. *Journal of Computational Chemistry*, 18:849–873, 1997.

[84] S. Miyazawa and R. L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules*, 18:534–552, 1985.

[85] A.G. Murzina, Steven E. Brennera, Tim Hubbarda, and Cyrus Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 7 April 1995.

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

[87] A. Neumaier. Molecular modeling of proteins and mathematical prediction of protein structure. *SIAM Rev.*, 39:407–460, 1997.

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

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

[90] 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, Birkhauser, Boston, MA*, pages 435–508, 1994.

[91] C.A Orengo, A. D. Michie, D.T. Jones, M.B. Swindells, and J.M.Thornton. CATH: A hierarchic classification of protein domain structures. *Structure*, pages 1093–110, 1997.

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

[93] I. Peterson. Simulations nab protein-folding mistakes. *Science News*, 155, 1999.

[94] G. N. Reeke. Protein folding: Computational approaches to an exponential-time problem. *Ann. Rev. Comput. Sci.*, 3:59–84, 1988.

[95] A. Sali, E. Shakhnovich, and M. Karplus. How does a protin fold? *Nature*, 369:248–251, 1994.

[96] W. R. Schmitt and M. Waterman. Linear trees and rna secondary structure. *Discrete Applied Mathematics*, 51:317–323, 1994.

[97] R. Schwartz, S. Istrail, and J. King. Frequencies of amino-acid strings in globular proteins sequences indicate suppression of blocks of consecutive hydrophobic residues. *Protein Science*, 10:1023–1031, 2001.

[98] E. Shakhnovich. Modeling protein folding: the beauty and power of simplicity. *Folding and Design*, 1:50–54, 1996.

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

[100] N. Sloane. Kepler's conjecture conffirmed. *Nature*, 395:435–436, 1998.

[101] P. D. Thomas and K. A. Dill. Statistical poteintials extracted from protein structures: How accurate are they? *J. Mol. Biol.*, 257:457–469, 1966.

[102] 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.*, 55:1183–1198, 1993.

[103] J. v. Neumann. The mathematician. *in "The Works of the Mind"*, University of Chicago Press:180–196, 1947.

[104] J. v. Neumann. Method in physical sciences. *in "The Unity of Knowledge" ed. L. Leary*, pages 157–164, 1955.

[105] J. v. Neumann and O. Morgenstern. Game theory and economic behaviour. *Princeton University Press*, 1944.

[106] W W. Kabash. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.*, page A32:922923, 1978.

[107] M. Wagner, J. Meller, and R. Elbert. Large-scale linear programming techniques for the design of protein folding potentials. *Math. Program. Ser. B*, 101:301–318, 2004.

[108] M. Waterman. Combinatorics of rna hairpins and cloverleaves. *Studies in Appl. Math.*, 60:91–96, 1978.

[109] M. Waterman and T. Smith. Rna secondary structure: A complete mathematical analysis. *Mathematical Biosciences*, 42:257–266, 1978.

[110] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai. Navigating the folding routes. *Science*, 267:1619, 1995.

[111] Y. Q. Zhou and M. Karplus. Interpreting the folding kinetics of helical proteins. *Nature*, 401:400–403, 1999.