JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 4, Number 1, 1997
Mary Ann Liebert, Inc.
Pp. 1–22

Robust Proofs of NP-Hardness for Protein Folding: General Lattices and Energy Potentials

WILLIAM E. HART and SORIN ISTRAIL

ABSTRACT

This paper addresses the robustness of intractability arguments for simplified models of protein folding that use lattices to discretize the space of conformations that a protein can assume. We present two generalized NP-hardness results. The first concerns the intractability of protein folding independent of the lattice used to define the discrete protein-folding model. We consider a previously studied model and prove that for *any* reasonable lattice the protein-structure prediction problem is NP-hard. The second hardness result concerns the intractability of protein folding for a class of energy formulas that contains a broad range of mean force potentials whose form is similar to commonly used pair potentials (e.g., the Lennard-Jones potential). We prove that protein-structure prediction is NP-hard for any energy formula in this class. These are the first robust intractability results that identify sources of computational complexity of protein-structure prediction that transcend particular problem formulations.

Key words: protein folding; intractability, robustness, lattice models.

1. INTRODUCTION

A PROTEIN IS A CHAIN OF AMINO ACID RESIDUES that folds into a unique *native* three-dimensional structure under physiological conditions. The problem of protein-structure prediction is a notoriously difficult problem in biochemistry. Proteins unfold when folding conditions provided by the environment are disrupted, and many proteins spontaneously refold to their native structures when physiological conditions are restored. This is the basis for the belief that prediction of the native structure of a protein can be done *computationally* from the information contained in the amino acid sequence.

Exhaustive search of a protein's conformational space is clearly not a feasible algorithmic strategy. The number of possible conformations is exponential in the length of the protein sequence, and powerful computational hardware would not be capable of searching this space for even moderately large proteins. This observation led Levinthal (Ngo *et al.*, 1994) to raise a question about the paradoxical discrepancy between the enormous number of possible conformations and the fact that most proteins fold within seconds to minutes, independent of their size. While these observations appear contradictory, they can be

Sandia National Laboratories, Algorithms and Discrete Mathematics Department, P.O. Box 5800, Albuquerque, New Mexico 87185-1110.

reconciled by noting that they simply point to the lack of knowledge of a possible algorithmic structure that could guide an efficient search algorithm (see Ngo *et al.,* 1994 for further discussion of this issue). Consequently, computational analyses of the protein folding process can provide insight into the inherent algorithmic difficulty of folding proteins.

Following the thermodynamic hypothesis (Epstein *et al.,* 1963), computational models of protein folding are typically formulated to find the global minimum of a potential energy function. Many simple protein-folding models use lattices to describe the space of conformations that proteins can assume. Lattices are infinite periodic graphs that are generated by translations of a "unit graph" that fill a two- or three-dimensional space. Lattices provide a natural discretization of the space of protein conformations since the dihedral angles along the protein's backbone are indeed constrained to specific domains. The conformation of a protein is often viewed as a self-avoiding path in the lattice in which the vertices are labeled by the amino acids (Dill *et al.,* 1995). An energy value is associated with every conformation taking into account neighborhood relationships of the amino acids on the lattice.

In this paper we explore the possible computational intractability of the problem of protein-structure prediction using techniques from computational complexity theory. Specifically, we use the theory of NP-completeness (Garey and Johnson, 1979). No polynomial algorithm has been constructed for any NP-complete problem. In fact, it is widely believed that no such polynomial-time algorithm exists.

Several lattice models of protein folding have been proven to be NP-hard (Ngo and Marks, 1992; Fraenkel, 1993; Unger and Moult, 1993; Patterson and Przytycka, 1995; Hart and Istrail, 1996), which means that they are at least as hard to solve as NP-complete problems. While these results support an algorithmic interpretation Levinthal's paradox, an important criticism of these results can be articulated as follows: "What is the biological relevance of a complexity analysis of structure prediction in *one* lattice model?" The relationship between previous complexity analyses is unclear because they consider different abstractions of the protein-folding problem. For example, authors have previously examined the complexity of lattice models on the three-dimensional cubic lattice (Fraenkel, 1993; Hart and Istrail, 1996; Patterson and Przytycka, 1995), on a "nearly-cubic" lattice (Unger and Moult, 1993), and on a diamond lattice (Ngo and Marks, 1992). The question we raise addresses the extent to which these complexity analyses might be specific to the particular details of these lattice models.

Results that transcend specific problem formulations are of significant interest because they may say something about the general biological problem with a higher degree of confidence. In fact, it is reasonable to expect that there will exist sources of computational complexity across lattice models that fundamentally relate to the protein-folding problem, since different lattice models provide discretizations of the same physical phenomenon. However, the identification of these sources of complexity has not been previously addressed.

In computational terms, the independence of algorithmic results from particular settings is called *computational robustness.* Robust algorithmic results are particularly important in computational models of protein folding, since complexity results for this class of optimization problems should be interpreted with caution. Accurate formulas for potential energy functions are not known; various analytic formulations use empirical potentials that attempt to represent the dominant physical forces (Creighton, 1993; van Gunsteren *et al.,* 1993). Learning from the computational complexity analysis of other optimization problems, we know that altering the problem objective even slightly (e.g., adding a one to the objective function) could change the status of a problem from NP-complete to tractable. Therefore, a robust analysis has a better chance of identifying sources of computational difficulty.

This paper presents robust complexity analyses for two lattice models that are related to the model described by Unger and Moult (1993). First, we analyze the computational complexity of Unger and Moult's model on an arbitrary three-dimensional lattice. Our analysis shows that the protein structure prediction problem is NP-hard for *any* reasonable lattice. Second, we analyze a restricted version of this problem on a cubic lattice. Our analysis examines the complexity of protein folding for a broad class of energy formula that are similar to commonly used pair potentials. We prove that protein-structure prediction is NP-hard for a class of energy formula for which the energy monotonically increases to zero with the distance between amino acids.

Definitions: Let $\mathbf{Z}$ be the set of integers, $\mathbf{Q}$ the set of rationals, $\mathbf{R}$ the set of reals. Let $\mathbf{Z}^+$ be the set of positive integers and $Z^{\geq 0}$ be the set of nonnegative integers. A vector v by convention has components $(v_1, \ldots, v_n)$. ■

2. LATTICE PROTEIN-FOLDING MODELS

A lattice protein-folding model represents conformations of proteins as vertex-independent embeddings of the protein structure in a lattice (i.e., no two amino acids are mapped to the same vertex). Lattice models can be classified based on the following properties:

1. The *physical structure,* which specifies the level of detail at which the protein's conformation is represented. The structure of the protein is treated as a graph whose vertices represent components of the protein. For example, we can represent a protein with a linear-chain structure (Dill *et al.,* 1995) that uses a chain of beads to represent the amino acids. Similarly, we can represent a protein with a simple side-chain structure (Bromberg and Dill, 1994) that uses a chain of beads to represent the backbone; amino acids are represented by edges to beads that connect to the backbone.
2. The *alphabet* of types of amino acids that are modeled by the problem. For example, we could use the 20 naturally occurring types of amino acids, or a binary alphabet that categorizes amino acids as hydrophobic (nonpolar) or hydrophilic (polar).
3. The *energy formula* used, which specifies how pairs of amino acid residues are used to compute the energy of a conformation. For example, this includes contact potentials that only have energy between amino acids that are adjacent on the lattice, and distance-based potential that use a function of the distance between points on the lattice. Many energy formulas have *energy parameters* that can be set to different values to capture different aspects of the protein folding process.
4. The *lattice,* in which protein conformations are expressed; this determines the space of possible conformations for a given protein. For example, the cubic and diamond lattices have been used to describe protein conformations (see Fig. 1).

A *conformation* of a protein sequence in an embedding of the protein's physical structure (i.e., the protein's graph) into the lattice such that vertices are mapped one-to-one to lattice points, and edges in the structure are mapped to corresponding lattice edges. We disallow a conformation to use two edges in a lattice that intersect, even though these edges may be included in the lattice.

The following graph-theoretic definition of lattices captures properties that are needed in our analysis [for a more conventional definition, see Wells (1979)]. Let $B = \{b_1, b_2, b_3\}$ be a set of linearly independent vectors in $\mathbf{R}^3$ (i.e., B spans $\mathbf{R}^3$). A *translation* of a point $p = (p_1, p_2, p_3)$ is a point defined by

$$\rho_1 b_1 + \rho_2 b_2 + \rho_3 b_3 + p,$$

where $(\rho_1, \rho_2, \rho_3) \in \mathbf{Q}^3$. A *primitive translation* is a translation for which $(\rho_1, \rho_2, \rho_3) \in \mathbf{Z}^3$.

We denote a *unit cell* by (V, η), where V is a set of points and $\eta : V \to \mathbf{R}^3$ is an embedding of these points in $\mathbf{R}^3$. We assume that V is finite. A lattice is generated by a unit cell (V, η) if all points on the lattice can be defined by a primitive translation of the points $\eta(v), v \in V$. A *unit graph* is a unit cell (V, η) with two types of edges: (a) $E \subseteq V \times V$, and (b) $E' \subseteq V \times V'$, where V' is a finite set of vertices outside of the unit cell. Thus, a unit graph is defined by a graph G and an embedding η', where (a) $G = (V \cup V', E \cup E')$ and (b) $\forall v \in V, \eta'(v) = \eta(v)$. We call the vertices in V *interior* vertices and

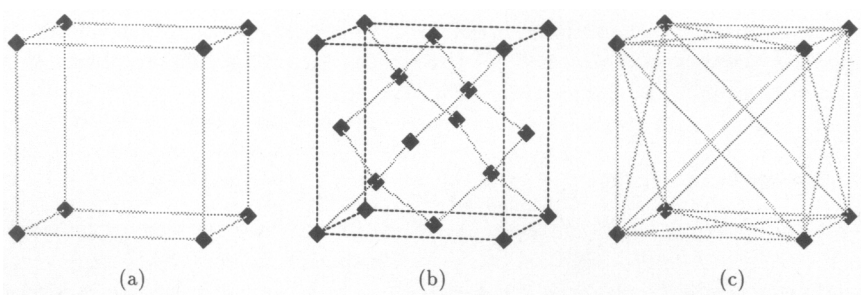

FIG. 1. Examples of unit graphs for lattices: (a) three-dimensional cubic, (b) diamond, (c) three-dimensional square with planar diagonal edges.

the vertices in V' *exterior* vertices. Examples of unit graphs are shown in Figure 1, using the conventional representation of the unit cell as a cube for which points on the faces of the cube are "shared" among neighboring unit cells.

Let B be a basis and A a unit graph, $A = (G, \eta)$, $G = (V \cup V', E \cup E')$. We say that A is *consistent* with B if there exists a primitive translation of A, $\overline{A}$, with the following property: for all $e \in E'$, if $\overline{A} = (\overline{G}, \overline{\eta})$, $\overline{G} = (\overline{V} \cup \overline{V}', \overline{E} \cup \overline{E}')$, then there exists $\overline{e} \in \overline{E}'$ such that $\eta(e) = \overline{\eta}(\overline{e})$. Consistency guarantees that connectivity between unit graphs is symmetric. If there is an edge in E' from A to $\overline{A}$, then there must be a complementary edge in $\overline{E}'$ from $\overline{A}$ to A.

Let $\mathcal{G}$ be an infinite periodic graph generated by translations of a unit graph. $\mathcal{G}$ is *connected* if there exists a path between any two vertices in it. Consider the graph H derived from $\mathcal{G}$ in which vertices represent unit graphs and edges represent the fact that two unit graphs share an edge. We say that $\mathcal{G}$ is connected if H is connected, which is a property common to physical lattices. A *lattice,* $\mathbf{L}$, is a connected infinite periodic graph generated by primitive translations with a basis B of a unit graph that is consistent with B. Note that the primitive translations of the unit graph "fill" all of $\mathbf{R}^3$ because B spans $\mathbf{R}^3$. We say that a lattice $\mathbf{L}$ is *finitely representable* if $b_i \in \mathbf{Q}^3$ and the coordinates of all points in the unit graph that defines $\mathbf{L}$ are vectors in $\mathbf{Q}^3$.

3. COMPUTATIONAL COMPLEXITY AND PROTEIN FOLDING

The *native* conformation of a protein is the conformation that has biological function. According to the thermodynamic hypothesis, the native conformation of a protein is the conformation with the minimum free energy among the set of all conformations. Consequently, given a lattice model (using lattice $\mathbf{L}$) and sequence s, the *protein-folding structure-prediction problem* (PFSP) is to find a native conformation of s in $\mathbf{L}$ with minimal energy.

Computational intractability refers to our inability to construct efficient (i.e., polynomial time) algorithms that can solve a given problem. Here, "inability" refers to both the present state-of-the-art of algorithmic research as well as possible mathematical statement that no such algorithms exist. Customary statements about the intractability of a problem are made by showing that the problem is NP-complete. The theory of NP-completeness provides overwhelming evidence towards the inexistence of polynomial time algorithms for NP-complete problems; the best known algorithm for any NP-complete problem takes an exponential number of computational steps, which makes these problems "practically intractable."

The class of problems NP includes a wide variety of notoriously difficult combinatorial problems, such as the traveling salesman problem, scheduling problems, and network design. Problems in NP have the property that given an instance of the problem and a potential solution, one can efficiently test to determine whether the potential solution actually solves the problem instance. A problem is NP-complete if it belongs to NP, and if there is a polynomial algorithm that can solve this problem, then this algorithm can be adapted to solve all of the other problems in NP. Hence, the problem is at least as hard as every other problem in NP. For a thorough treatment of NP-completeness see Garey and Johnson (1979).

Formally, NP-complete problems are decisions problems, for which the answer is either yes or no. Optimization problems like PFSP are not directly considered within the framework of NP-completeness. However, optimization problems can be transformed into a decision problem by introducing a threshold B and asking whether a solution with value less than or equal to B exists. The corresponding optimization problem is at least as hard as the decision problem, since finding the optimal solution would answer this decision problem for every value of B. Consequently, an optimization problem is NP-hard if its corresponding decision problem is shown to be NP-complete.

3.1. Previous NP-hardness results

Several authors use this framework to prove that PFSP is NP-hard for various lattice protein-folding models. Fraenkel (1993) examines a physical model in which each amino acid is represented as a bead in a graph. The graph represents the contacts in the protein that must be held at a fixed distance, presumably including the edges along the backbone of the protein. The alphabet consists of three types that represent the charges associated with the amino acids: $-1, 0, 1$. The model uses a distance-dependent energy formula

that computes the product of the charges divided by distance. The energy is the sum over all edges in the contact graph that is provided in the problem specification. A cubic lattice is used with this model.

Ngo and Marks (1992) present a hardness result for a molecular structure prediction problem that encompasses protein folding. This model considers a chain molecule of atoms whose energy is based upon a typical form of the empirical potential-energy function for organic molecules. Conformations of this chain molecule are embedded in a diamond lattice.

Paterson and Przytycka (1995) examine a physical model in which each amino acid is represented as a bead along a chain. Their model allows for an unbounded number of different types of amino acids. A contact energy formula is used, which has contact energies of one for contacts between identical residues and zero otherwise. The lattice used by this model is the cubic lattice.

Finally, Unger and Moult (1993) examine a protein-folding model that applies to the lattice defined by the unit graph in Figure 1c. Their model treats amino acids as beads along a chain. The energy formula is a simple form of a free energy function that has the same form as empirically derived force fields (Unger and Moult, 1993). Hart and Istrail (1996) generalize this NP-hardness result to Bravais lattices (which includes the cubic lattice), as well as the diamond and fluorite lattices.

3.2. Robust notions of intractability

It is difficult to provide strong recommendations for particular protein-folding models because accurate potential energy functions are not known. While various analytic formulations use potentials that capture known features of "the" potential function, the most appropriate analytic formulation of the potential energy for protein folding remains an area of active research (Creighton, 1993; Gunsteren *et al.*, 1993). Consequently, robust algorithmic results are particularly important for computational models of protein folding.

Computational robustness refers to the independence of algorithmic results from particular settings. In the context of NP-completeness, robustness refers to the fact that a class of closely related problems can be described, all of which are NP-complete. The members of the class of problems are typically distinguished by some parameter(s) that form a set of reasonable alternate formulations of the same basic problem. For example, we can define robustness with respect to the lattice used in a particular lattice model. In this case, each member of this class of problems is defined with respect to a particular lattice.

In the next two sections, we describe robustness arguments for two different classes of protein-folding models. First, we describe an intractability result that is robust to changes in the lattice. We examine the model proposed by Unger and Moult (1993) and demonstrate that this model remains NP-hard for any reasonable lattice. Next, we describe an intractability result that is robust to changes in the energy formulas. We examine a class of energy formulas that captures a wide range of potentials of mean force that monotonically increase to zero as the distance between amino acids increases.

4. A HARDNESS RESULT FOR GENERAL LATTICES

4.1. Model formulation

Consider the following lattice protein-folding model. The physical structure specifies that the protein sequence $S = s_1, \ldots, s_n$ is treated as an n-vertex node-labeled path, where node i is labeled with $s_i, i = 1, \ldots, n$. Each node on the path represents a single amino acid in the protein. The alphabet of amino-acid types are represented by integers, $\mathcal{A} = \{1, \ldots, m\}$. Here, $m \leq n$, but the value of m may depend upon n, so the alphabet size is not bounded above by a constant value.

Let $F_S = \{f_1, \ldots, f_n\}$ represent a conformation of S, where $f_i \in \mathbf{Q}^3$ is the position of the amino acid s_i in $\mathbf{L}$. A conformation F_S is an embedding of S in $\mathbf{L}$ where every f_i is at a vertex of $\mathbf{L}$ and between the vertices f_i and f_{i+1} there exists an edge in $\mathbf{L}$. Furthermore, an embedding is not a valid conformation if it contains two edges that cross. Suppose that $f_i = (x_i, y_i, z_i)$. Then $d_x(f_i, f_j) = |x_i - x_j|$, $d_y(f_i, f_j) = |y_i - y_j|$, and $d_z(f_i, f_j) = |z_i - z_j|$. The energy formula for this model is

$$\sum_{i=2}^{n} \sum_{j=1}^{i-1} C_{s_i, s_j} g(d_x(f_i, f_j), d_y(f_i, f_j), d_z(f_i, f_j)),$$

where C is an m by m matrix and and $g : \mathbf{Q}^3 \to \mathbf{R}^{\geq 0}$ is a symmetric function.[2] Both C and g are energy parameters that depend upon the particular instance of this model that is being considered. The function g provides a simple form of a free energy function that is appropriate to the lattice approximation. Furthermore, the form of this energy formula is similar to empirically derived force fields (Unger and Moult, 1993).

To analyze the computational complexity of this problem, we cast it into a decision problem, **L**-PF:

Instance: A sequence $S = (s_1, \ldots, s_n), s_i \in \{1, \ldots, m\}$ such that $m \leq n$; a symmetric nonnegative function $g : \mathbf{Q}^3 \to \mathbf{R}^{\geq 0}$ with finite representation; a matrix $C \in Z^{m \times m}$; $B \in \mathbf{Z}$.

Question: Is there a conformation F_S embedded in **L** such that

$$\sum_{i=2}^{n} \sum_{j=1}^{i-1} C_{s_i, s_j} g(d_x(f_i, f_j), d_y(f_i, f_j), d_z(f_i, f_j)) \leq B?$$

Note that this problem is lattice-specific, so we have defined a class of decision problems.

4.2. *Results and discussion*

Theorem 1 shows that **L**-PF is NP-complete for any finitely representable lattice **L**. This model is a slight variant of the model proposed by Unger and Moult (1993). Specifically, we have restricted our analysis to energy formulas for which the function g is symmetric. We believe that this restriction makes our hardness argument more physical. Because g is a symmetric function, it captures some of the translational invariance that is true of natural potential energy functions. Since instances of our model form a subset of the instances of Unger and Moult's model, this hardness result implies that their model for PFSP is NP-hard for any lattice.

Theorem 1. *Let* **L** *be a finitely representable lattice. Then* **L**-PF *is NP-complete.*

This hardness result provides the first evidence that the difficulty of protein folding is not simply an artifact of the particular discretization used to formulate the PFSP problem. Such discretizations are a natural step to enable the application of the tools of computational complexity to this problem. This result suggests that arguments that use computational complexity to show that PFSP is difficult may be robust to changes in the lattice formulation that they use.

Furthermore, this result temper's criticism of NP-hardness results for lattice models because they're not continuous (Ngo *et al.*, 1994). Because this result applies for every lattice, the precision required by the user can be supplied through an arbitrarily refined lattice. Since all practical protein-folding algorithms must work to a given precision, these lattice formulations reflect the type of algorithmic implementations that would be applied by a practical algorithm.

4.3. *Technical results*

When a unit graph is translated in each dimension, each vertex v in the unit graph has corresponding vertices in each of the translated copies of the unit graph. We distinguish these vertices by labelling them with integral coordinates $v_B(i, j, k)$, where (i, j, k) denotes the primitive translation of v with respect to B; we omit the reference to B when it is clear from context. We will also use these coordinates to refer to the unit graphs themselves.

Before describing the proof of Theorem 1, we describe how a unit graph can be constructed for a lattice such that the set of interior vertices in the unit cell are located within a cubic volume with integral extent (i.e., a cube with integer-length edges). Recall that e_i is the unit vector $(0, \ldots, 0, 1, 0, \ldots, 0)$, where the 1 is in the ith dimension.

[2]We say that a function is *symmetric* if its value does not change when its inputs are reordered (e.g., $g(x, y, z) = g(z, x, y)$).

Lemma 1. *If* $\mathbf{L}$ *is a finitely representable lattice generated with a unit graph* C *and basis* B*, then there exists a unit graph* C' *that can generate* $\mathbf{L}$ *using the basis* $\overline{B} = \{Ke_1, Ke_2, Ke_3\}$, $K \in \mathbf{Z}^+$*, such that (a) the interior vertices of* C' *lie strictly within a cubic volume with extent* K *in each dimension and (b) the exterior vertices of* C' *lie strictly outside this cubic volume.*

Proof. Let $B = \{b_1, b_2, b_3\}$ be the basis used to define $\mathbf{L}$. Let the unit graph for $\mathbf{L}$ be $C = (G, \eta)$, $G = (V \cup V', E \cup E')$, and let $p = \eta(v(0, 0, 0))$ for some $v \in V$.

We begin by demonstrating that the standard Cartesian basis $\{e_1, e_2, e_3\}$ can always be used to define the primitive translations of points on the lattice. To do this, we show that there exists a $\kappa_1 \in \mathbf{Z}^{\geq 0}$ such that $p + \kappa_1 e_1$ is a vertex in $\mathbf{L}$. Let p' be this new point. Because p' is generated via translations according to B, we have

$$p' = p + \rho_1 b_1 + \rho_2 b_2 + \rho_3 b_3 = p + \kappa_1 e_1$$

If $b_i = (b_{i,1}, b_{i,2}, b_{i,3})$, then we can expand this to get

$$\begin{bmatrix} \rho_1 b_{1,1} + \rho_2 b_{2,1} + \rho_3 b_{3,1} \\ \rho_1 b_{1,2} + \rho_2 b_{2,2} + \rho_3 b_{3,2} \\ \rho_1 b_{1,3} + \rho_2 b_{2,3} + \rho_3 b_{3,3} \end{bmatrix} = \begin{bmatrix} \kappa_1 \\ 0 \\ 0 \end{bmatrix}.$$

Consider the last two of these linear equations. There clearly exists a set of values $\rho' = (\rho'_1, \rho'_2, \rho'_3)$, $\rho'_i \in \mathbf{Q}$, that satisfies these linear equations such that ρ' is not the vector of all zeros. Let $\kappa' = \rho'_1 b_{1,1} + \rho'_2 b_{2,1} + \rho'_3 b_{3,1}$. Note that $\kappa' \neq 0$, since otherwise the b_i would be linearly dependent. In general, κ' may be a fractional value. Suppose that $\kappa' = \alpha/\beta$. Then the vector $(\beta\rho'_1, \beta\rho'_2, \beta\rho'_3)$ satisfies the last two of the linear equations and $\kappa_1 = \beta\rho'_1 b_{1,1} + \beta\rho'_2 b_{2,1} + \beta\rho'_3 b_{3,1}$ is a nonzero integer.

A similar argument shows that there exist integers κ_2 and κ_3 that can be used to define primitive translations along e_2 and e_3. These translations can be used to redefine $\mathbf{L}$ using the basis $B' = \{\kappa_1 e_1, \kappa_2 e_2, \kappa_3 e_3\}$ and a unit graph C' defined by the convex hull of the eight vertices $v_B(\overline{\kappa})$, $\overline{\kappa} \in \{0, 1\}^3$, where the interior vertices of C' are those contained within this volume as well as those contained on the "bottom," "front" and "left" faces of the volume and the exterior vertices are all other vertices to which the interior vertices are connected.

In general, $\kappa_i \neq \kappa_j$ for $i \neq j$, so the volume that contains C' is a parallelepiped. Let $K = \kappa_1\kappa_2\kappa_3$. Now consider the basis $\overline{B} = \{(K/\kappa_1)e_1, (K/\kappa_2)e_2, (K/\kappa_3)e_3\}$ and the volume defined by the eight vertices $v_{\overline{B}}(w)$, $w \in \{0, 1\}^3$. This unit volume is composed of repetitions of volume that contains C' such that C' is repeated K/κ_i times along the ith dimension. Because the extent of C' along the ith dimension is κ_i, the extent of this new volume along the ith dimension is K, which is integral. Now we can translate this cubic volume along the vector $(1, 1, \ldots, 1)$ by an arbitrarily small amount to ensure that there are no vertices that lie on any of its faces. We use the translated cubic volume to define our final unit graph C'. Vertices within the translated cubic volume are the interior vertices of C' and the exterior vertices are all other vertices to which the interior vertices are connected. ■

The following lemma extends Lemma 1 to show that there not only exists a unit graph contained within a cubic volume with integral extent, but that a unit graph exists for which each unit graph is connected by an edge to each of its six neighboring unit graphs. This connectivity between neighboring unit graphs enables us to consider paths through the lattice in a manner analogous to the paths constructed on the 3D cubic lattice. The following assumption describes the conditions that Lemma 2 guarantees can be satisfied by a unit graph of $\mathbf{L}$.

Assumption 1. Consider a unit graph for which

1. the interior vertices of the unit graph are strictly within a cubic volume with length $K \in \mathbf{Z}^+$ in all dimensions,
2. there exists edges of the unit graph that are connected to each of the six neighboring unit graphs.

Lemma 2. *If* $\mathbf{L}$ *is a finitely representable lattice generated with a unit graph* C *and basis* B*, then there exists a unit graph* C' *satisfying Assumption 1 that can generate* $\mathbf{L}$ *using the basis* $\overline{B} = \{Ke_1, Ke_2, Ke_3\}$, $K \in \mathbf{Z}^+$.

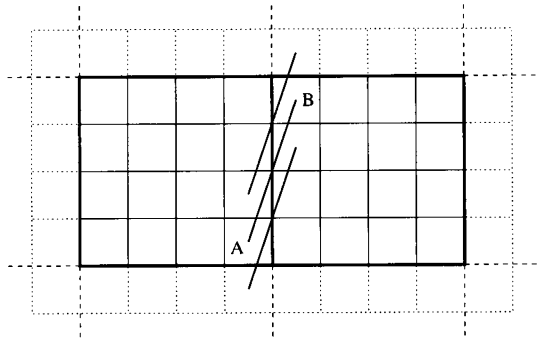

FIG. 2. Illustration of the edges connecting neighboring unit graphs. The edge $\overline{AB}$ is guaranteed to exist because the extent of the unit graph is scaled with J_y.

Proof. We know from Lemma 1 that the first condition of Assumption 1 can be satisfied. Suppose that $\overline{C}$ is a unit graph that satisfies this condition. Consider the edges that connect the $(0, 0, 0)$th unit graphs to other unit graphs. Let J_x be the largest integer (in absolute value) such that the $(0, 0, 0)$th unit graph is connected to the (J_x, a, b)th unit graph (for some integers a and b). From the connectivity property of lattices we know that there exists such an integer. We similarly define J_y and J_z to be the closest connections in the y- and z-dimensions. Now consider a unit volume with extent $J_x J_y J_z K$ whose corners are the cubic volumes of the eight unit graphs with indeces in $\{0, J_x J_y J_z K\}^3$. This cubic volume can be used to define a unit graph C' that satisfies both conditions of Assumption 1. Because the unit graph has been expanded by a factor that is at least as great as the length of any edge along each dimension, there always exists an instance of $\overline{C}$ on the face of C' that is connected to an instance of the neighbor of C'. Figure 2 illustrates this in two-dimensions. ■

The following corollary simply notes that the unit graph that satisfies Assumption 1 can be quickly constructed from a unit graph and a basis.

Corollary 1. *Given a unit graph C and a basis B, a unit graph satisfying Assumption 1 can be constructed in a number of steps that is polynomial in the size of C.*

The proof of Theorem 1 uses a reduction from the Optimal Linear Arrangement problem (OLA) (Garey and Johnson, 1979):

Instance: A graph $G = (V, E)$; a positive integer B.

Question: Is there a one-to-one function $f : V \to \{1, 2, \ldots, |V|\}$ such that

$$\sum_{\{u,v\}\in E} |f(u) - f(v)| \leq B?$$

Proof of Theorem 1. From Corollary 1 we know that a unit cell for $\mathbf{L}$ that satisfies Assumption 1 can be constructed in polynomial time. In the remainder of the proof we assume that such a unit cell has been constructed.

Let v be a vertex within the unit cell. Let p_y^i be the shortest path from $v(i, j, k)$ to $v(i, j+1, k)$ in which all vertices are either in the (i, j, k)th or $(i, j+1, k)$th unit cell. We know from Assumption 1 that such a path always exists. Now p_y^j and p_y^{j+1} may not be vertex disjoint, so we need to identify vertices between which the shortest paths are vertex disjoint. Let the symmetric difference of p_y^j and p_y^{j+1} be defined by the path whose edges are the symmetric difference of the edges in p_y^j and the edges in p_y^{j+1}. Consider the intersection of p_y^j and p_y^{j+1}. This path intersects the symmetric difference of p_y^j and p_y^{j+1} at a single vertex, v_y (see Fig. 3). Let $\hat{p}_y^j$ be the shortest path between $v_y(i, j, k)$ and $v_y(i, j+1, k)$. Now $\hat{p}_y^j$ and $\hat{p}_y^{j+1}$ are vertex disjoint, so we can construct a path along the y dimension using the v_y vertices. Let D_y be the length of p_y^i. We define v_x^j, v_z^j, D_x and D_z similarly. Let $D = D_x D_y D_z$.

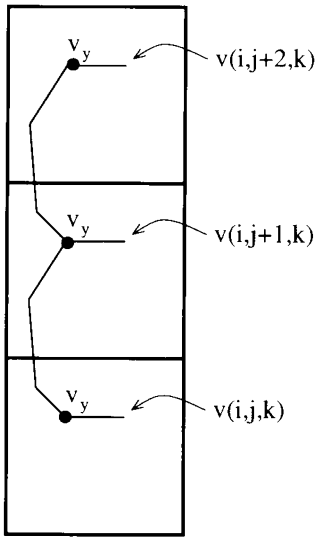

FIG. 3. Illustration of the identification of the vertices v_y.

Let a y^+x^- turn be the symmetric difference of the shortest path from $v_y(i, j, k)$ to $v(i, j+1, k)$ and the shortest path from $v(i, j+1, k)$ to $v_x(i-1, j+1, k)$. The y^+x^- turn is a turn that goes through a unit cell, entering from a path connected to a vertex v_y along the y-axis and exiting along a path connected to a vertex v_x along the x-axis (see Fig. 4). In this fashion, we define the following types of turns: (1) y^+x^-, (2) x^-z^-, (3) z^-x^+, (4) x^+y^-, (5) y^-x^+, (6) x^+z^-, (7) z^-x^-, and (8) x^-y^+. We will use these numbers to refer to these turns.

Now consider a y^+x^- turn from $v_y(i, j, k)$ to $v_x(i-1, j+1, k)$. The length of a y^+x^- turn is approximately equal to $D_y + D_x$. We define T_1 such that the length of a y^+x^- turn is $D_y + D_x + T_1$ (T_1 can be either negative or positive). Here, the "1" refers to the index of the turn in the list above. The values T_i are similarly defined for the other turns, $i = 2, \ldots, 8$.

We are now prepared to describe how an instance of OLA can be transformed to L-PF. Let $m = |V| + 1$ and let $a_i = i, i = 1, \ldots, m-1$, be the amino acids that correspond to the vertices in V. Let $x = m$ be the label for the remaining amino acid type. Consider

$$S = a_1 \underbrace{xxx \ldots xx}_{\kappa_O} a_2 \underbrace{xxx \ldots xx}_{\kappa_3} a_3 \underbrace{xxx \ldots xx}_{\kappa_O} a_4 \ldots a_n,$$

where $\kappa_O = D(8m) - (T_1 + T_2 + T_3 + T_4)$ and $\kappa_e = D(8m) - (T_5 + T_6 + T_7 + T_8)$. The costs are

$$C_{s_i,s_j} = \begin{cases} |f(s_i) - f(s_j)| & \text{if } s_i, s_j \in \{a_1, \ldots, a_{m-1}\} \\ 0 & \text{otherwise} \end{cases},$$

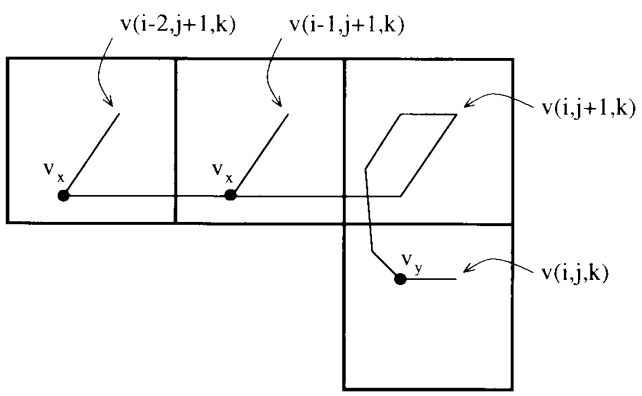

FIG. 4. Illustration of a y^+x^- turn.

We use the same parameter B to bound the energy as in the OLA instance. Let $\xi = d_x[v(i, j, k), v(i + 1, j, k)]$. For any pair of vertices v_1 and v_2, if there is a value ν such that the distance between v_1 and v_2 is $2\xi\nu$ along the x, y and z dimensions, then the function g is ν. Otherwise, g takes a very large value. This penalizes interactions between vertices that are not along a line parallel to the vector (1, 1, 1) and which are not spaced evenly apart along this line. Formally, if $\exists\nu$ such that

$$d_x(v_1, v_2) = d_y(v_1, v_2) = d_z(v_1, v_2) = 2\xi\nu$$

then

$$g(d_x(v_1, v_2), d_y(v_1, v_2), d_z(v_1, v_2)) = \nu.$$

Otherwise,

$$g(d_x(v_1, v_2), d_y(v_1, v_2), d_z(v_1, v_2)) = (B + 1)/C_{\min},$$

where $C_{\min}$ is the smallest nonzero cost in C.

Small energies are only possible if the a_i lie along a line in the three-dimensional lattice that is parallel to the vector (1, 1, 1). Furthermore, in the optimal conformation, the a_i must be separated by an odd number of unit cells along this line. If a conformation violates these constraints, its energy will be at least $B + 1$ and the solution will be rejected.

To show that the answer to L-PF is yes exactly when the answer to OLA is yes, we show that (1) every possible conformation for the original OLA problem has a corresponding conformation in the L-PF problem, and (2) the accepted solution to L-PF corresponds to an accepted solution to OLA. Figure 5 illustrates the type of conformations used to generate any ordering of a_i via vertex disjoint paths. This figure does not show the precise path taken on the lattice, but instead describes the unit cells within which the path lies.

Since D is a multiple of D_x, D_y and D_z, a path can always be constructed that passes through vertices v_x, v_y and v_z. We use this observation as follows. Starting at a vertex v_y, the a_i are connected using paths on different planes of unit cells. Residues a_i and a_{i+1} are connected on plane $(2m - 3 + i)D$ for odd i or plane $-iD$ for even i. The chain alternately moves vertically up and down through the column of unit cells containing the a_i, using the paths on the different planes to form loops between the a_i. These paths are vertex disjoint because all vertical paths lie within different columns of unit cells, and the horizontal paths forming the loops lie on different planes of unit cells.

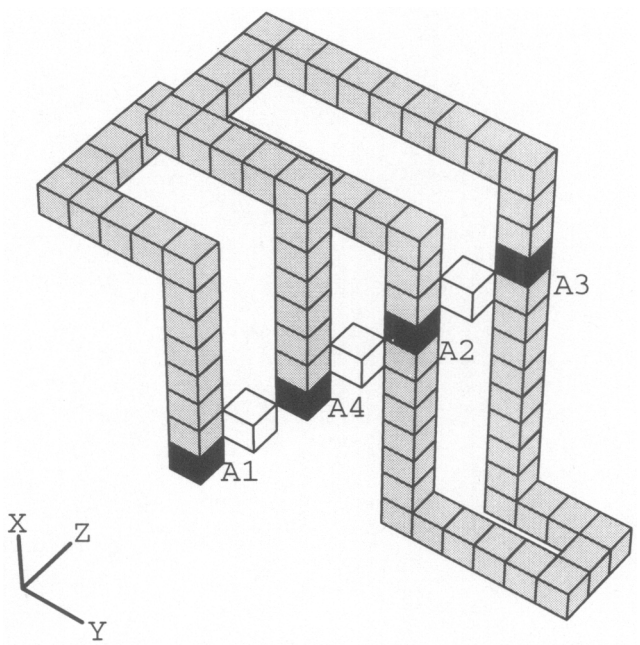

FIG. 5. Illustration of the conformational invariant needed to transform OLA to L-PF. Gray cubes are unit cells that contain only x amino acids and black cubes are unit cells that contain an a_i amino acid. White cubes are used to illustrate the diagonal line on which the black cubes are located. This conformation is an example of a transformation used when $D = 1$.

The lengths of the sequences of xs are sufficient to allow any pair of a_i to be connected in the optimal conformation. In the worst case this conformation needs to connect vertices $v_y(i, j, k)$ and $v_y(i + 2(m - 2), j + 2(m - 2), k + 2(m - 2))$ via a path on plane $-(m - 2)D$ or $(3m - 5)D$. If the path is via the plane $(3m - 5)D$, then its length is $D(8m) - (T_1 + T_2 + T_3 + T_4)$. Otherwise, the path's length is $D(8m) - (T_5 + T_6 + T_7 + T_8)$. The differences between these lengths accounts for the different turns used between pairs of a_i when i is odd and when i is even.

For the sequence generated by the reduction, only the interactions between the a_i contribute to the conformational energy, which is the same as in the OLA problem. To see that any accepted solution for **L**-PF corresponds to an accepting solution in OLA, simply observe that any **L**-PF solution that does not have the a_i lie along a line parallel to $(1, 1, 1)$ at even multiples has an energy of at least $B + 1$. Consequently, such solutions are not accepted. Any other accepted solution must have the elements along the line with an energy below $B + 1$, so it is an accepted solution of OLA.

Since the expansion of the original OLA problem is polynomial, it follows that **L**-PF is NP-hard. Checking the energy of a conformation to verify that it is below $B + 1$ is polynomial, so **L**-PF is NP-complete. ■

5. A HARDNESS RESULT FOR GENERAL ENERGY FORMULAS

In this section, we examine the complexity of protein folding for a class of protein-folding models that encompasses a broad range of energy formulas. The following assumption defines conditions on a function that restrict it to be monotonically increasing towards zero. This class of functions represents the pairwise potentials of mean force that we consider, where the domain of the function is the distance between two amino acids. These functions do not include the repulsive component of the commonly used pair potentials. However, the lattice implicitly enforces an excluded volume constraint that accounts for certain types of repulsive forces.

Assumption 2. Consider a function $g : \mathbf{Q} \to \mathbf{R}$ such that

1. $\forall x \in \mathbf{Q}, x \geq 1, g(x) \leq 0$
2. $\forall x, y \in \mathbf{Q}$, if $x > y$ then either $g(x) > g(y)$ or $g(x) = g(y) = 0$.
3. $\exists C_1 \in \mathbf{Q}^+$, $\exists C_2 \in \mathbf{Q}^{\geq 0}$ and $\exists p_1, p_2 \in \mathbf{Q}^+$, $p_1 > 1$, $p_2 > 1$, such that $\forall x \in \mathbf{Q}^+, x \geq 1, -C_1/x^{p_1} \leq g(x) \leq -C_2/x^{p_2}$.

One of the most studied pairwise potential functions is the Lennard-Jones potential (Allen and Tildesley, 1987). This is a simple idealized pairwise potential commonly used in computer simulations. It reflects the salient features of atomic interactions in a general, often empirical way. The so called Lennard-Jones 12-6 potential is:

$$\lambda_{\epsilon,\sigma}(x) = 4\epsilon\left[\left(\frac{\sigma}{x}\right)^{12} - \left(\frac{\sigma}{x}\right)^{6}\right].$$

It is easy to see that $\lambda_{\epsilon,\sigma}$ belongs to our class of functions.

5.1. *Model formulation*

Consider the following lattice protein-folding model. The physical model specifies that the protein sequence $S = s_1, \ldots, s_n$ is treated as an n-vertex node-labeled path, where node i is labeled with $s_i, i = 1, \ldots, n$. Each node on the path represents a single amino acid in the protein. The alphabet of amino-acid types are represented by integers, $\mathcal{A} = \{1, \ldots, m\}$. Here, $m \leq n$, but it may depend upon n, so the alphabet size is not bounded by a constant value.

Let $F_S = \{f_1, \ldots, f_n\}$ represent a conformation of S, where $f_i \in \mathbf{Q}^3$. A conformation F_S is an embedding of S in the cubic lattice if every f_i is in $\mathbf{Z}^3$ and for all i there exists an edge between the vertices f_i and f_{i+1} (i.e., they are neighbors on the lattice). The energy formula for this model is

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1} C_{s_i,s_j} g(d(f_i, f_j)),$$

where C is an m by m matrix, $g : \mathbf{Q} \to \mathbf{R}$ is a function that satisfies Assumption 2, and $d : \mathbf{Q}^3 \times \mathbf{Q}^3 \to \mathbf{R}$ is a distance measure.

This model can be viewed as a special case of the model examined by Unger and Moult (1993). The restrictions that we place on the energy formula preclude criticisms of their work (Ngo *et al.*, 1994) concerning the lack of translational invariance that can occur in specific instances of their objective function. An energy formula is translationally invariant if the energy of a protein is independent of its location and orientation in space. The model we propose is translationally invariant because g uses only the value returned by d, so the energy of a protein conformation remains the same in any orientation.

To analyze the computational complexity of this problem, we cast it into a decision problem, (g, d)-**PF**:

Instance: A sequence $S = (s_1, \ldots, s_n)$, $s_i \in \{1, \ldots, m\}$ such that $m \leq n$; a matrix $C \in \mathbf{Z}^{m \times m}$; $B \in \mathbf{Q}$.

Question: Is there a conformation F_S embedded in the cubic lattice such that

$$\sum_{i=2}^{n} \sum_{j=1}^{i-1} C_{s_i, s_j} g(d(f_i, f_j)) \leq B?$$

Note that this problem depends upon the definitions of g and d, so we have defined a class of decision problems; members of this class of problems correspond to particular choices of g and d.

5.2. *Results and discussion*

Theorem 2 shows that (g, d)-PF is NP-complete for any function g that satisfies Assumption 2 and for d functions that are discretized versions of the p-norm. Recall that the L_1 norm measures the length of a vector $v = (v_1, \ldots, v_n)$ and $|v|_1 = \sum_{i=1}^{n} |v_i|$. Similarly, the L_2 norm measures the length of a vector v as $|v|_2 = \sqrt{\sum_{i=1}^{n} v_i^2}$. We define a L_2^i norm to be the value of the L_2 norm with i decimal values of precision: $\lfloor |v|_2 * 10^i \rfloor / 10^i$.

Theorem 2. *Let g be a function that satisfies Assumption 2 and let d be either the L_1 norm or the L_2^i norm, $i \geq 1$. Then* (g, d)-PF *is NP-complete.*

This result can be generalized to a broader range of problems in two ways. First, Assumption 2 can be weakened to include functions that are bounded by a function of the form $-C/h(x)^p$, where $h(x) > 0$ for all $x \in \mathbf{Q}^+$, $x \geq 1$ and h is efficiently invertible. For example, Assumption 2 can be weakened to include functions of the form $-C/(\log(x + 1))^p$ because logarithms are easily inverted.

Second, this result can be generalized to include a broader range of distance metrics. The L_1 norm and the L_2^i norm ($i \geq 1$) both have the property that the distance from a point on the cubic lattice to one of its six nearest neighbors is one, while the distances to all other points on the lattice is greater than one. This property is used in our analysis to enforce the construction of particular conformational structures. This property can also be true for L_p^j norms, which compute the L_p norm to j decimal values of precision:

$$\left\lfloor \left(\sum_{i=1}^{n} \right)^{1/p} * 10^j \right\rfloor / 10^j.$$

As p increases, the value of j needed to insure the property described above increases. However, for any finite value of p, the requisite value of j value is finite and so there is a class of (g, d)-PF problems defined by these L_p^j norms d that are NP-complete.

5.3. *Technical details*

We prove Theorem 2 using a reduction from HAMILTONIAN PATH (Garey and Johnson, 1979):

Instance: A graph $G = (V, E)$.

Question: Does there exist a Hamiltonian path in G?

A *Hamiltonian path* is a path that passes through every vertex in the graph such that each vertex is traversed exactly once.

We prove Theorem 2 by showing that HAMILTONIAN PATH reduces to (g, d)-PF. In our reduction, it is relatively easy to demonstrate that if there exists a Hamiltonian path then the protein sequence generated

in our reduction can be configured into a low-energy state. To prove the converse we need to show that lowest energy conformation of the protein sequence is a unique structure that can be interpreted to represent a Hamiltonian path if one exists.

The following lemmas illustrate how subsequences of the protein generated in our reduction form lowest energy structures. Lemma 3 shows how a parallelepiped structure can be constructed. Subsequent lemmas show how the protein sequence used in Lemma 3 can be extended to form conformational structures that build upon the parallelepiped.

Lemma 3. *Let $N \in \mathbf{Z}^+$, $N \geq 3$. Consider the sequence*

$$P = r_1 \ldots r_N s_N \ldots s_1 t_1 \ldots t_N T_N \ldots T_1 S_1 \ldots S_N R_N \ldots R_1,$$

where the r_i, s_i, t_i, R_i, S_i, and T_i are $6N$ distinct amino acid types. Let $\Delta < 0$, and let the cost matrix C be defined as follows:

- $C_{r_i,s_i} = \Delta, i = 1, \ldots, N-1$
- $C_{s_i,t_i} = \Delta, i = 2, \ldots, N$
- $C_{S_i,T_i} = \Delta, i = 2, \ldots, N$
- $C_{R_i,S_i} = \Delta, i = 1, \ldots, N-1$
- $C_{r_i,R_i} = \Delta, i = 1, \ldots, N$
- $C_{s_i,S_i} = \Delta, i = 1, \ldots, N$
- $C_{t_i,T_i} = \Delta, i = 1, \ldots, N-1$
- $C_{i,j} = 0$ otherwise

Then the unique lowest energy structure for this protein sequence is a $3 \times N \times 2$ parallelepiped (see in Fig. 6).

Proof. Note that the parallelepiped illustrated in Figure 6 has $7N - 5$ contacts. Because each amino acid has a distinct set of interactions with the other amino acids, this is exactly the maximal number of contacts with energy Δ that a conformation of P can make. Except for obvious symmetries, this structure is unique because of the constraints imposed by the contacts themselves. There is only one configuration of the contacts between $r_1, r_2, R_1, R_2, s_1, s_2, S_1$, and S_2. Given this, if we inductively consider additional amino acids in P, there remains a single configuration with lowest energy. ■

The following lemma extends the previous result to determine the location of other amino acids on the surface of the parallelepiped. We assume for the moment that we can ignore the remainder of the protein sequence that connects these amino acids together as well as the subsequence that connects them to the parallelepiped. It will be convenient to let $\bar{s}_i$ refer to the amino acid $s_{2(i-1)\delta+2}$ for some $\delta \in \mathbf{Z}^+$.

Lemma 4. *Let $N = 2(n-1)\delta + 3$ for some $\delta \in \mathbf{Z}^+$. Consider the sequence P defined in Lemma 3. Let $a_i, i = 1, \ldots, n$, be n distinct amino acids. We extend the cost matrix C by adding interactions $C_{a_j,s_i} = 1$, for $i = 2, 2\delta + 2, \ldots, 2(n-1)\delta + 2$ and $j = 1, \ldots, n$. If Δ is sufficiently negative and δ is sufficiently large, then the unique lowest energy structure for this protein sequence places the a_i along the face of the parallelepiped (see Fig. 7).*

Proof. Since our main interest is in the configuration of the a_i on the parallelepiped, we simply assume that Δ is sufficiently negative that the energy for a single contact in the parallelepiped is more negative than the total energy of the interactions between the $\bar{s}_i$ and a_j, regardless how the a_j and $\bar{s}_i$ are configured. This ensures that the parallelepiped exists in the lowest energy configuration.

We will show that when a sufficiently large value of δ is selected, the lowest energy configuration contains n contacts between the a_j and the $\bar{s}_i$. We begin by showing that the contact energy between amino acid a_j and $\bar{s}_i$ is more negative than the total sum of all of the interactions between a_j and the other $\bar{s}_k, k \neq i$, if a_j and $\bar{s}_i$ do not form a contact.

To force a contact between a_j and $\bar{s}_i$, we need to have

$$g(1) < \min_{f \in H} \sum_{k=1}^{n} g(d(f, f_{\bar{s}_k})),$$

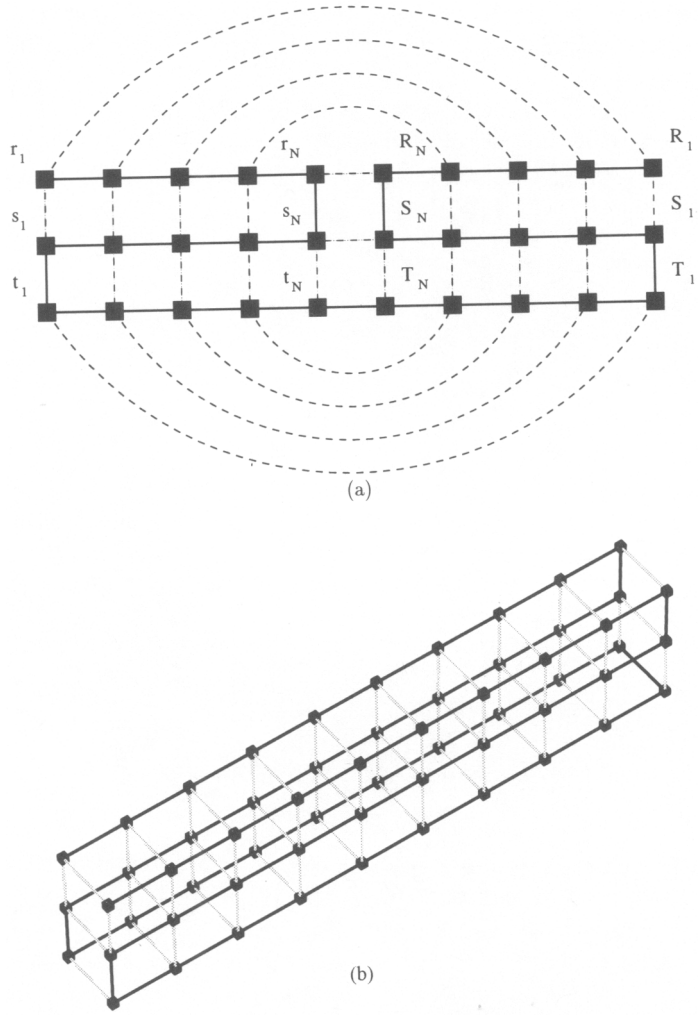

FIG. 6. Illustration of the lowest energy conformational configuration assumed by the protein sequence defined in Lemma 3: (a) the two halves of the parallelepiped (dashed lines represent contacts, which have energy Δ) and (b) the three-dimensional structure. Dark edges represent the conformation of the protein instance. Gray and dashed edges represent contacts between the amino acids.

where H is the set of possible locations where a_j could be placed that do not form a contact with a $\bar{s}_k$, and $f_{\bar{s}_k}$ is the location of $\bar{s}_k$. Now suppose that there exists $f \in H$ such that

$$g(d(f_{a_j}, f_{\bar{s}_i})) = g(1) \geq \sum_{k=1}^{n} g(d(f, f_{\bar{s}_k})).$$

Let $d_1, \ldots, d_n$ be the distances $d(f, f_{\bar{s}_k})$ sorted into ascending order. Clearly, $d_1 \geq 2$ since f_{a_j} cannot represent a contact with a $\bar{s}_k$. Now both $d_2 \geq \delta$ and $d_3 \geq \delta$ since otherwise $d_1 > \delta$. Similarly, we have $d_{2k} \geq \delta(2k-1)$ and $d_{2k+1} \geq \delta(2k-1)$. Since g is monotonically increasingly, we have

$$\begin{aligned}
\sum_{k=1}^{n} g(d(f, f_{\bar{s}_k})) &\geq g(2) + \sum_{k=1}^{\lfloor n/2 \rfloor} 2g(\delta(2k-1)) \\
&\geq g(2) + \sum_{k=1}^{\lfloor n/2 \rfloor} \frac{-2C}{\delta^p (2k-1)^p} \\
&\geq g(2) - 2\lfloor n/2 \rfloor C/\delta^p,
\end{aligned}$$

since $g(x) \geq -C/x^p$.

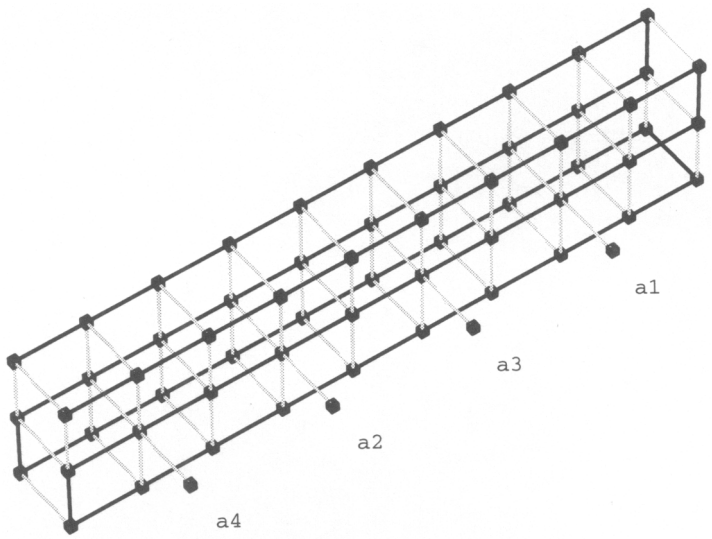

FIG. 7. Illustration of the lowest energy conformational configuration assumed by the protein sequence defined in Lemma 3 along with the n amino acids $a_i(\delta = 1)$. Dark edges represent the conformation of the protein instance. Gray edges represent contacts between the amino acids.

To form a contradiction, it suffices to show that

$$g(1) < g(2) - 2\lfloor n/2 \rfloor C/\delta^p.$$

This can be achieved by choosing δ such that

$$\delta \geq \left\lceil \left(\frac{nC}{g(2) - g(1)} \right)^{1/p} \right\rceil + 1.$$

Because we have forced a contradiction, we know that a_j is forced to form a contact with $\bar{s}_i$ for sufficiently large δ, since violating this contact increases the energy more than the sum of the interactions of a_j with all of the $\bar{s}_k$. This argument applies to each of the a_j independently. Consequently each of the a_j must form a contact with one of the $\bar{s}_k$. ■

In the final lemma, we describe a protein sequence that has enough flexibility to form any permutation of contacts of the a_j amino acids with the $\bar{s}_k$.

Lemma 5. *Let $N = 2(n-1)\delta + 3$ for some $\delta \in \mathbf{Z}^+$ and let $\bar{n} = 2(n+N)+1$. Consider the sequence*

$$S = r_1 \ldots r_N s_N \ldots s_1 t_1 \ldots t_N T_N \ldots T_1 S_1 \ldots S_N R_N \ldots$$
$$R_1 \underbrace{xxx\ldots xx}_{\bar{n}} a_1 \underbrace{xxx\ldots xx}_{\bar{n}} a_2 \underbrace{xxx\ldots xx}_{\bar{n}} \ldots \underbrace{xxx\ldots xx}_{\bar{n}} a_n,$$

where x is an amino acid type that has zero for its contact interaction with all amino acids. If Δ is sufficiently negative and δ is a sufficiently large number, then the unique lowest energy structure for this protein sequence forms some permutation of contacts between the a_i and the $\bar{s}_k$ along the face of the parallelepiped (see Fig. 8).

Proof. Suppose that Δ and δ are given as described by the proof of Lemma 4. The a_i are connected using paths on different planes. Residues a_i and a_{i+1} are connected on plane $-i-2$ for odd i or plane $i+2$ for even i. The chain alternately moves vertically up and down through the column containing the a_i, using the paths on the different planes to form loops between the a_i. The lengths of the sequences of xs are more than sufficient to allow any pair of a_i to be connected. R_1 and a_1 are connected using plane 2. Further, the number of xs between a_i and a_{i+1} is odd, which permits these two amino acids to be placed at an even distance apart. ■

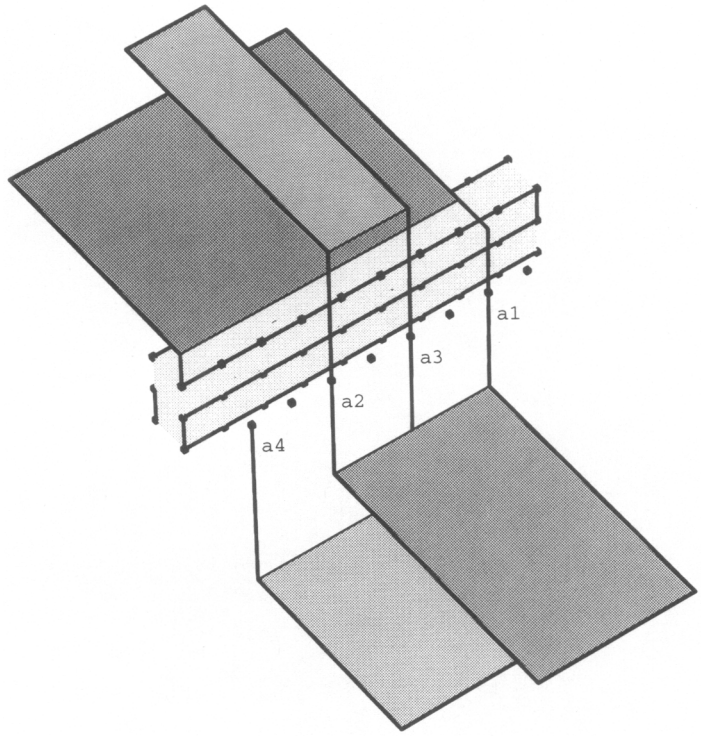

FIG. 8. Illustration of the lowest energy conformational configuration assumed by the protein sequence defined in Lemma 5.

We now present the proof of Theorem 2.

Proof. Let $C_1 \in \mathbf{Q}^+$, $C_2 \in \mathbf{Q}^{\geq 0}$ and $p_1, p_2 \in \mathbf{Q}^+$, $p_1 > 1$, $p_2 > 1$, such that $\forall x \in \mathbf{Q}^+, x \geq 1, -C_1/x^{p_1} \leq g(x) \leq -C_2/x^{p_2}$. Let

$$\delta = \max\left\{\left\lceil\left(\frac{nC}{g(2)-g(1)}\right)^{1/p}\right\rceil, \left\lceil\left(\frac{C_1 n^2}{C_2 3^{p_1}}\right)^{1/(p_1-p_2)}\right\rceil\right\} + 1,$$

$W = g(1)$, and $N = (2n-1)\delta + 3$.

To transform an instance of HAMILTONIAN PATH to (g,d)-PF, we construct a protein instance as follows. Let A be a set of amino acid types that includes

- a_i that correspond to the vertices in V
- y_i that are related to the vertices in V
- a "dummy" amino acid x
- $6N$ other amino acid types:

$$r_1, \ldots, r_N, s_1, \ldots, s_N, t_1, \ldots, t_N, R_1, \ldots, R_N, S_1, \ldots, S_N, T_1, \ldots, T_N.$$

Let $\bar{n} = 2(n+N)+1$. We construct a sequence

$$\begin{aligned} S = {} & r_1 \ldots r_N s_N \ldots s_1 t_1 \ldots t_N T_N \ldots T_1 S_1 \ldots S_N R_N \ldots \\ & R_1 \underbrace{xxx\ldots xx}_{\bar{n}} a_1 \underbrace{xxx\ldots xx}_{\bar{n}} a_2 \underbrace{xxx\ldots xx}_{\bar{n}} \ldots \underbrace{xxx\ldots xx}_{\bar{n}} a_n xx \\ & y_n \underbrace{xxx\ldots xx}_{\bar{n}} y_{n-1} \underbrace{xxx\ldots xx}_{\bar{n}} \ldots \underbrace{xxx\ldots xx}_{\bar{n}} y_1. \end{aligned}$$

Let

$$\Omega = \left\lceil \frac{-n^2 W}{g(2)-g(1)} \right\rceil + 1$$

and

$$\Delta = \left\lceil \frac{-3n^2\Omega W}{g(2) - g(1)} \right\rceil + 1.$$

The cost matrix C is defined as follows:

- $C_{r_i,s_i} = \Delta, i = 1, \ldots, N-1$
- $C_{s_i,t_i} = \Delta, i = 2, \ldots, N$
- $C_{S_i,T_i} = \Delta, i = 2, \ldots, N$
- $C_{R_i,S_i} = \Delta, i = 1, \ldots, N-1$
- $C_{r_i,R_i} = \Delta, i = 1, \ldots, N$
- $C_{s_i,S_i} = \Delta, i = 1, \ldots, N$
- $C_{t_i,T_i} = \Delta, i = 1, \ldots, N-1$
- $C_{s_j,a_i} = \Omega, i = 1, \ldots, n; j = 2, 2\delta+2, \ldots, 2(n-1)\delta+2$
- $C_{s_j,y_i} = \Omega, i = 1, \ldots, n; j = \delta+2, 3\delta+2, \ldots, (2n-1)\delta+2$
- $C_{a_j,y_j} = 1, j = 1, \ldots, n$
- $C_{a_j,y_i} = 1, \{j,i\} \in E$
- $C_{i,j} = 0$, otherwise

Let $J = (7N-5)g(1)$ and let

$$H = \sum_{i=1}^{n}\sum_{j=1}^{n} g(d(\{2(i-1)\delta+2, 1, 0\}, \{2(j-1)\delta+2, 0, 0\})).$$

H represents the total interaction between the a_i and s_j when the a_i form contacts at s_j, $j = 2, 2\delta+2, \ldots, 2(n-1)\delta+2$. We define the threshold $B = J\Delta + 2H\Omega + (2n-1)g(\delta)$.

Given an instance of HAMILTONIAN PATH, the configuration shown in Figure 9 provides very low energy for the first part of the protein sequence (up until the first x). These contacts have a total energy of $J\Delta$. Furthermore, Figure 9 shows the locations where the a and y amino acids make contacts with this configuration. If $v_{i_1} v_{i_2} \ldots v_{i_n}$ is the Hamiltonian path in the graph, then we can construct a conformation that lays the a and y amino acids at these contact points in the order $a_{i_1} y_{i_1} a_{i_2} y_{i_2} \ldots a_{i_n} y_{i_n}$. This conformation lays the xs in alternating planes as shown in Figure 10. This is similar to the conformation described by Lemma 5. The basic idea is to fold the sequence up until a_n as is illustrated in Figure 8, and then thread the remaining sequence in the reverse order, shifted along the parallelepiped and using unique planes to loop the xs. Each of the a_i interact with each of the s_j, $j = 2, 2\delta+2, \ldots, 2(n-1)\delta+2$. The energy of these

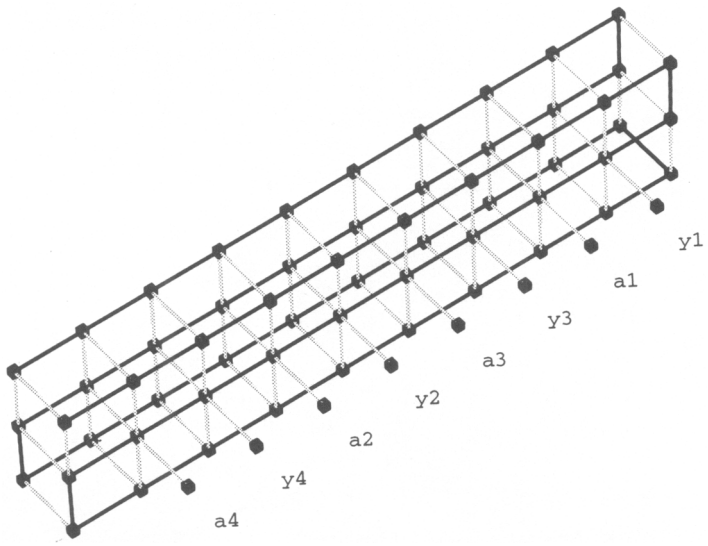

FIG. 9. Illustration of the conformational configuration assumed by the dominant energetic interactions in the transformed protein instance. Dark edges represent the conformation of the first part of the protein instance. Gray edges represent contacts between the amino acids in this configuration.

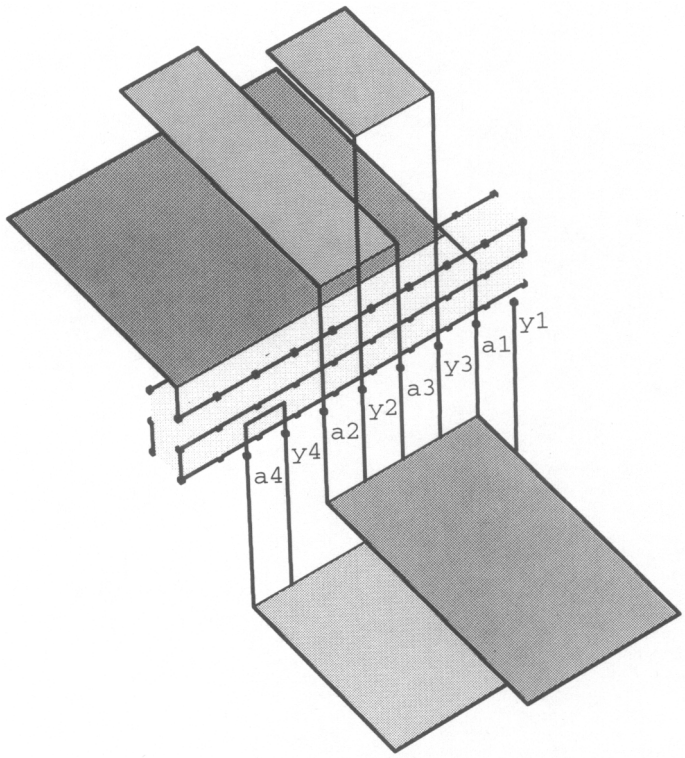

FIG. 10. Illustration of a native conformation in the transformed protein instance. Gray planes highlight the different planes on which the xs are looped. To clarify the total structure of the conformation, the conformation assumed by the first part of the protein sequence (up until the first x) is placed in a light gray box.

interactions is $H\Omega$. Similarly, the interactions between y_i and s_j is $H\Omega$. Finally, each a_i interacts with the neighboring y_i and each y_i interacts with its neighboring a_j. Thus the interactions between the y_i and a_j have an energy less than $(2n-1)g(\delta)$. Thus the total energy is less than $J\Delta + 2H\Omega + (2n-1)g(\delta) = B$, as desired.

Now suppose the protein sequence is in a conformation with energy less than or equal to B. We show that the unique lowest energy conformation of the protein sequence is the structure shown in Figure 10 in three steps. First, we show that the r, s, t, R, S, and T amino acids must form the contacts shown in Figure 9. There are n^2 interactions between the a_i and s_i, which have a total energy that is no smaller than $n^2\Omega W$. Similarly, the interactions between the y_i and s_i have a total energy that is no smaller than $n^2\Omega W$. There are at most n^2 interactions between the a_i and y_i, which have a total energy of n^2W or more. Thus the total energy of the interactions that are not interactions between the r, s, t, R, S, and T amino acids is greater than or equal to

$$2n^2\Omega W + n^2 W > 3n^2\Omega W.$$

If a single contact amongst the r, s, t, R, S, and T amino acids is not made, then the total energy of the protein's conformation is greater than

$$(7N-4)\Delta W + \Delta g(2) + 3n^2\Omega W.$$

Now

$$\Delta = \left\lceil \frac{-3n^2\Omega W}{g(2) - g(1)} \right\rceil + 1,$$

from which it follows that $\Delta g(2) + 3n^2\Omega W > \Delta g(1)$. Consequently, if a single contact amongst the r, s, t, R, S, and T amino acids is not made, then the total energy is greater than B, which is a contraction. Lemma 3 shows that if all of the contacts amongst the r, s, t, R, S, and T amino acids are made, then the lowest energy conformation of this section of the protein sequence is the parallelepiped shown in Figure 9.

Next, we demonstrate that the a_i and y_j form contacts with the s_k, as shown in Figure 9. There are two independent reasons why these contacts might not be formed: (1) violating a contact might lead to a conformation with lower total energy between the a_i or y_j and s_k, and (2) violating one or more contacts might lead to a conformation with lower energy between the a_i and y_j. From Lemma 4 we know that because $\delta \geq \left\lceil \left(\frac{nC}{g(2)-g(1)} \right)^{1/p} \right\rceil + 1$ that the first case is not true. If a single contact between the a_i or y_j and s_k is violated, then the conformational energy is no less than

$$\Delta J + 2H\Omega - \Omega g(1) + \Omega g(2) + n^2 W.$$

Since

$$\Omega = \left\lceil \frac{-n^2 W}{g(2) - g(1)} \right\rceil + 1,$$

we know that $-\Omega g(1) + \Omega g(2) + n^2 W > 0$. This implies that the conformational energy is greater than B, which is a contradiction.

Finally, we need to show that the a_i and y_j form a sequence along the parallelepiped such that neighboring a_i and y_j add $g(\delta)$ to the total energy. Suppose that there exists neighbors a_i and y_j such that $C_{a_i,y_j} = 0$. Then the total energy of the interactions between the a_i and y_j is greater than or equal to $(2n-2)g(\delta) + \sum_{i=2}^{n} \sum_{j=1}^{i-1} g((2(i-j)+1)\delta)$. We have $\delta \geq \left\lceil \left(\frac{C_1 n^2}{C_2 3^{p_1}} \right)^{1/(p_1-p_2)} \right\rceil + 1$, from which we have

$$g(\delta) \leq g(3\delta)n^2 < \sum_{i=2}^{n} \sum_{j=1}^{i-1} g((2(i-j)+1)\delta).$$

This implies that the total energy of the interactions is greater than $(2n-1)g(\delta)$, which contradicts the assumption that the total energy of the protein's conformation is greater than B. Because the y_i bridge the gaps between subsequent a_j only when the a_j have an edge between them in G, the sequence of a_j represents a Hamiltonian path in G. ■

6. GENERALIZATIONS

In this section we describe how the intractability analysis in the previous section can be extended to apply to a protein-folding model that explicitly represents side chains. It is possible that our intractability results depend upon the fact that the protein is represented as a linear chain. Our intractability analysis for the side-chain model demonstrates that the robustness argument for generalized energy potentials is not limited by this simplification.

The side-chain model that we consider was proposed by Bromberg and Dill (1994). In this model, the backbone is represented by a linear chain and the side chains are represented by vertices that are connected to each of the backbone vertices. We assume that the only energetic interactions are between the side-chain vertices. The formal description of this model is analogous to the description of (g, d)-PF, and we call this problem (g, d)-SCPF. Let

$$S = r_1 U_1 r_2 U_2 \ldots r_N U_N x x x R_N s_N \ldots R_1 s_1 x x x t_1 S_1 \ldots t_N S_N x x x T_N u_n \ldots T_1 u_1$$
$$\underbrace{xxx\ldots xx}_{\bar{n}} a_1 \underbrace{xxx\ldots xx}_{\bar{n}} a_2 \underbrace{xxx\ldots xx}_{\bar{n}} \ldots \underbrace{xxx\ldots xx}_{\bar{n}} a_n xx$$
$$y_n \underbrace{xxx\ldots xx}_{\bar{n}} y_{n-1} \underbrace{xxx\ldots xx}_{\bar{n}} \ldots \underbrace{xxx\ldots xx}_{\bar{n}} y_1,$$

where $N = n\delta + 1$ and $\bar{n}$ is much larger than N and n. Contact energies can be defined between r and R, s and S, t and T and u and U such that first part of S forms a parallelepiped as shown in Figure 11. The remainder of the sequence can permute the y_i and a_j just like the linear chain, to form a sequence along the face of the parallelepiped that represents a hamiltonean path in the graph is one exists. Consequently, the proof of NP-completeness for (g, d)-SCPF is analogous to the proof of NP-completeness for (g, d)-PF.

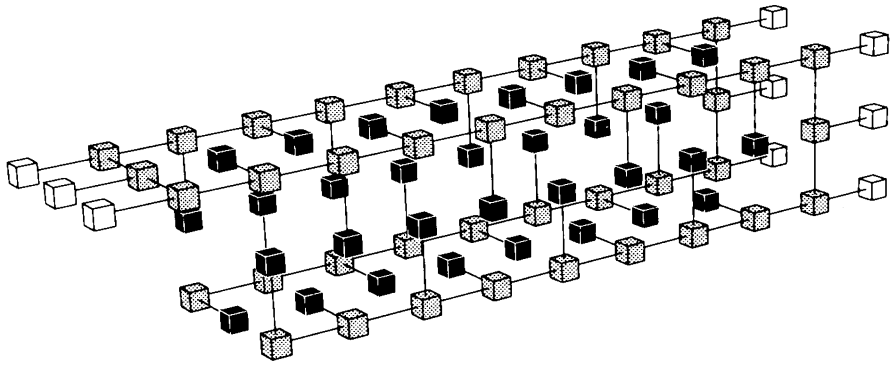

FIG. 11. Illustration of the parallelepiped constructed for the reduction to (g, d)-SCPF.

7. CONCLUSIONS

Several authors have provided detailed discussions of the relevance of intractability arguments to protein folding in natural systems (Fraenkel, 1993; Ngo and Marks, 1992; Ngo *et al.*, 1994; Unger and Moult, 1993). We review the major points raised by these authors and discuss the contribution of robust computational complexity arguments to our understanding of the biological protein-folding process.

A major factor that affects the interpretation of NP-hardness results is the extent to which the protein-folding model captures features of the protein-folding process that are fundamentally related to the time needed to perform protein folding. If the protein-folding models that have been analyzed *do* capture these features, then their associated NP-hardness results rule out the possibility that every amino acid sequence can be quickly folded to its native conformation (unless all NP-complete problems can be solved in polynomial time, which is highly unlikely). There are several interpretations that could reconcile this result with the fact that proteins reliably fold to their native state very quickly:

- Nature has selected only proteins for which the native conformation is attainable in polynomial time. Because NP-hardness is a worst-case concept, there may be a subset of amino acid sequences that can be folded in polynomial time by an algorithm. The NP-hardness simply rules out the possibility of this algorithm folding *all* proteins.
- The native conformation is not necessarily at a global minimum (Unger and Moult, 1993). This implies that the problem formulation does not accurately reflect the dynamics of protein folding, so there may be proteins that can only be quickly folded to native, low energy states that are not at the global minimum of the free energy formula.
- Nature can solve NP-hard problems in polynomial time. Fraenkel (1993) raises this possibility, though he does not propose an interpretation of the physical process that would indicate how this could be done for protein folding. One interpretation of this would be that the massive parallelism inherent in the relatively independent folding of solutions of proteins amounts to a brute force method of solving otherwise intractable problems. While this may be possible, it is unclear whether the length of proteins is within the range at which such a brute force folding to native states is possible.

These issues raise important questions about the nature of protein folding that need to be addressed, since they fundamentally relate to the nature of the process that we wish to analyze with the tools of computational complexity. Even if an NP-hardness result is constructed for a model that truly captures the difficulty of protein folding, we still need to determine whether our assumptions about the protein-folding process are well-modeled as a global energy minimization of all amino acid sequences.

If our models do *not* capture features of the protein-folding process that are fundamentally related to the time needed to perform protein folding, then NP-hardness results may simply reflect this fact. If the protein-folding model is too general, then the PFSP problem may be NP-hard simply because it contains instances that are not biologically plausible. An *instance* of the computational formulation of

the problem may contain parameters that vary both the protein sequence and energy formula. So for example, the model may contain instances where the free energy formula is not reasonable.

Because good models of protein folding have been difficult to construct (i.e., models whose free energy formula can accurately predict the structure of biologically relevant native conformations), this is an important consideration when evaluating complexity arguments. In fact, we argue that all protein-folding models for which PFSP has been proven NP-hard are too general in both of the ways that we have just described:

- Fraenkel's model (Fraenkel, 1993) permits constraint graphs that may not force subsequent amino acids to lie in close proximity on the lattice, thereby leading to biologically implausible native states for certain amino-acid sequences.
- Ngo and Marks' model (Ngo and Marks, 1992) is the most specific model that has been considered, since it contains the energy formulas commonly used to determine the shape of organic molecules. However, this formulation has a large number of parameters that are not constrained to lie within biologically plausible ranges. Furthermore, this energy formula does not incorporate compactness requirements that account for the effects of attractive Van der Waals forces and hydrophobicity (solvent entropy) (Ngo and Marks, 1992).
- Patterson and Przytycka's model (Patterson and Przytycka, 1995), while similar to other lattice protein-folding models (e.g., see Dill *et al.,* (1995)), does not restrict the number of different types of amino acids, thereby permitting an unbounded number of amino-acid types.
- The model examined by Unger and Moult (1993) and Hart and Istrail (1996) contains a very general energy formula. While the form of this free energy function is similar to that used for empirically derived mean force functions, its lack of specificity and potential for artificial energy instances is certainly a factor that makes PFSP difficult.[3]

This survey clearly shows that complexity arguments need to be developed for more specific models that are of particular interest. The robust intractability arguments that we have described represent an important step towards understanding the complexity of more interesting models. We have examined generalizations of models that include both general lattice formulations as well as general energy functions. The later analysis includes a class of energy formula that are commonly used to represent energy potentials, so this result shows that the computational complexity is not limited to specific energy formulations.

Because we have focused on a class of protein-folding problems instead of a specific problem, our results provide stronger evidence for the intractability of protein folding. However, our results do suffer from a weakness common to most of the protein-folding models discussed above. Specifically, the set of amino-acid types used to construct protein sequences is not bounded in size; there exist problem instances in the problems that we have analyzed for which no two amino acids can be categorized in the same class based upon their interactions with all of the other amino acids.

The consequence of this feature of our models is that the protein sequences do not have the property of *correlation* between amino acids (Chan and Dill, 1996). Chan and Dill (1996) argue that correlation is an important property of accurate physical models. At present, the model analyzed by Fraenkel (1993) is the only model with a finite number of amino-acid types whose complexity has been analyzed. This model is unsatisfactory, however, because it allows amino acids that are adjacent along the chain to lie arbitrarily far away on the lattice. An open problem raised by this work is whether PFSP is NP-hard for any model with a finite number of amino-acid types that embeds chains in a lattice in the manner that we have defined. For example, does there exist a lattice model using contact energies and a finite number of amino acids whose complexity is NP-hard? Since this type of model has been well studied (Dill *et al.,* 1995), this type of result could provide valuable insight into the computational aspects of PFSP.

[3]Unger and Moult (1993) argue that because their simple model is hard, solving more realistic models is even harder. However, this argument fails to recognize that more realistic models undoubtably contain constraints that could in principle make them easier to solve. In general, subclasses of NP-hard problems are not necessarily NP-hard.

ACKNOWLEDGMENTS

This work was supported by the MICS Applied Mathematical Sciences program, U.S. Department of Energy, Office of Energy Research, and was performed at Sandia National Laboratories, operated for the U.S. Department of Energy under contract No. DE-AC04-94AL85000. We thank Cindy Phillips, Phil MacKensie, and Sarina Bromberg for their helpful discussions.

REFERENCES

Allen, M.P., and Tildesley, D.J. 1987. *Computer Simulation of Liquids.* Oxford Science Publications, Oxford.

Atkins, J., and Hart, W.E. 1997 (Unpublished research).

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

Chan, H.S., and Dill, K.A. 1996. Comparing folding codes for proteins and polymers. *PROTEINS: Structure, Function, and Genetics* 24, 335–344.

Creighton, T.E., ed. 1993. *Protein Folding.* W.H. Freeman and Co., New York.

Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D., and Chan, H.S. 1995. Principles of protein folding: A perspective from simple exact models. *Prot. Sci.* 4, 561–602.

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

Fraenkel, A.S. 1993. Complexity of protein folding. *Bull. Math. Biol.* 55(6), 1199–1210.

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

Hart, W.E., and Istrail, S. 1996. Invariant patterns in crystal lattices: Implications for protein folding algorithms (extended abstract). *In Proc. 7th Annual Symp. on Combinatorial Pattern Matching,* 288–303.

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

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

Patterson, M., and Przytycka, T. 1995. On the complexity of string folding. *In* Istrail, S., Pevzner, P., and Shamir, R. eds. *Discrete Applied Mathematics, Special Issue on Computational Molecular Biology.* Vol. 71, No. 1–3, December 1996, pp. 217–230.

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

van Gunsteren, W.F., Weiner, P.K., and Wilkinson, A.J., eds. 1993. *Computer Simulation of Biomolecular Systems.* ESCOM Science Publishers, New York.

Wells, A.F. 1979. *Three-Dimensional Nets and Polyhedra.* American Crystallographic Association, New York.

Address reprint requests to:
William E. Hart
Sandia National Laboratories
Algorithms and Discrete Mathematics Department
P.O. Box 5800
Albuquerque, NM 87185-1110

wehart@cs.sandia.gov

Received for publication September 11, 1996; accepted as revised December 12, 1996.