Studies of protein designability using reduced models

by

Myron Peto

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Robert Jernigan, Co-major Professor
Drena Dobbs, Co-major Professor
David Fernandez-Baca
Mark Hargrove
Kai-Ming Ho
Oliver Eulenstein

Iowa State University

Ames, Iowa

2007

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank my current advisor, Dr. Jernigan, who graciously took me on mid-stream in my grad-school career. His lab has been a great fit for me and I appreciate all his support. I also should mention other lab mentors, including Andrzej Kloczkowski and Taner Sen. They helped get me started and provided encouragement and technical advice along the way.

The administrative team of the BCB program deserves mention, including Trish Stauble, Kathy Wiederin, Chris Tuggle, and Dan Voytas. During my five years here if the going got tough (which it did quite often) I knew I could go to them for a sympathetic ear. Kathy and Trish especially have been a big help in keeping my paperwork in order. Lord knows this is not one of my strengths.

Finally and most importantly, my family deserves a most hearty thanks. Tracie, my wife, moved from her Portland home to the Midwest in order for me to pursue this degree. This sacrifice on her part will not go forgotten. Both of my daughters, Eris and Sophie, have always presented a loving face to daddy and make coming home at the end of the day a welcome event.

ABSTRACT

One the most important problems in computational structural biology is protein designability, that is, why protein sequences are not random strings of amino acids but instead show regular patterns that encode protein structures. Many previous studies that have attempted to solve the problem have relied upon reduced models of proteins. In particular, the 2D square and the 3D cubic lattices together with reduced amino acid alphabets have been examined extensively and have lead to interesting results that shed some light on evolutionary relationship among proteins. Here, additionally to the 2D square lattice, we study the 2D triangular and 3D face centered cubic (fcc) lattices, we perform designability studies using different shapes embedded in the 2D square lattice, and we use machine learning algorithms to classify binary sequences folding to highly- or poorly-designable conformations.

In the first part of the thesis we extend the transfer matrix method to the 2D triangular lattice. The transfer matrix method is a highly efficient method of enumerating all conformations within a compact lattice area that has earlier been developed for the 2D square and 3D cubic lattices. In addition we also enumerated all compact conformations within simple geometries on the 2D triangular and 3D face centered cubic (fcc) lattices using a standard backtracking algorithm.

In the second part of the thesis we described protein designability studies on various shapes in the 2D square lattice using a reduced hydrophobic-polar (HP) amino acid alphabet. We used a simple energy function that counted the number of H-H, H-P and P-P interactions within a restricted set of protein shapes that have the same number of residues and non-bonded contacts. We found a difference in the designabilities of different protein shapes.

Finally, in the third part of the thesis we used standard machine learning algorithms to classify two classes of protein sequences. We first performed a designability study for two shapes, using a binary HP alphabet, on the 2D triangular lattice and separated highly- and poorly-designable conformations. Highly-designable conformations had many sequences folding to them with the lowest energy and poorly-designable conformations had few or no sequences folding to them. Sequences were classified as highly- or poorly-designable depending on whether they folded to highly- or poorly-designable structures. Using several machine learning algorithms such as Decision Tree, Naïve Bayes, and Support Vector Machine, we were able to classify highly- and poorly-designable sequences with high accuracy.

CHAPTER 1. INTRODUCTION

Within the field of structural biology there has been and is still considerable interest in application of simple reduced models to the study of protein structure, function and dynamics. These models span a wide range of complexity, from all-atom models with a realistic energy function to the simplest square lattice models with a reduced H/P binary amino acid alphabet. Within the body of work comprising this thesis we have focused on the simpler end of that spectrum – lattice models with a binary amino acid alphabet. There is much evidence that despite their simplicity the lattice models capture the essence of protein behavior. Lattice models have the significant advantage of requiring fewer conformations to fully sample their conformations, in contract to finer-grained all-atom models. Designability studies have shown that the vast majority of protein conformations are "non-designable" in the sense that few or no protein sequences would fold to them with energy lower than energies of all other conformations. In addition, designable conformations tend to show some of the features of real protein structures, such as symmetries of shape and structural flexibility.

Thesis Organization

The three main chapters cover studies using lattice models or extensions of tools developed for lattice models. Chapter 2 deals with the extension of the transfer matrix method, originally developed for the square and cubic lattices, to the 2D triangular lattice. Chapter 3 is a study of the impact of protein shape on protein designability using the 2D square lattice. Chapter 4 is an application of machine learning algorithms developed within the field of computer science to a study of protein designability on the 2D triangular lattice. Each of these chapters correspond to papers that have either been published or submitted.

Protein conformations are often modeled as self-avoiding walks on a pre-defined lattice area. Real globular protein structures tend to be densely packed, which is one reason why a compact self-avoiding walk without voids on a lattice area offers a reasonable representation. Many structural and functional studies using reduced models employ 2D square or 3D cubic lattices, having coordination numbers of 4 and 6, respectively. The body of the past work itself speaks of the utility and validity of these two simple lattices, but they do suffer from the parity shortcomings. Parity refers to the even/odd checkerboard nature of these lattices. Before any two points (residues) on a walk (protein chain) can be nearest neighbors on the lattice there must be an even number of points (residues) between them along the walk (chain). The 2D triangular and 3D face centered cubic (fcc) lattices do not suffer from the parity constraints. In addition, with coordination numbers of 6 and 12 for the 2D triangular and 3D fcc lattices, respectively, we believe that they offer a better representation of actual bond angles found in real protein structures.

We enumerate and generate all compact conformations (for paths and circuits) within numerous compact shapes embedded in the 2D triangular and 3D face centered cubic (*fcc*) lattices. These conformations are used to model protein structures and such complete generations of conformations allow an exhaustive search of the protein conformational space within the confines of a lattice area, something not possible for continuous models.

The transfer matrix method is a way of enumerating and generating all self-avoiding walks that fit within a defined area of a given lattice. Other brute force methods of generating all possible conformations suffer from the serious problem of attrition – with increasing chain length most walks reach a dead end without visiting all sites within a predefined shape on the lattice. Their method has previously been developed for the 2D square and 3D cubic lattices [1-4]. In addition, an application of the transfer matrix method using cooperative potentials to study the statistical averages of conformational ensembles has been developed [5].

In our work we have extended the transfer matrix method to the triangular 2D lattice, which allows us to efficiently generate conformations on a lattice of higher coordination that does not have the parity issue mentioned above. Because of the increased coordination number, the number of conformations within a given area, relative to the 2D square lattice, can be more representative even thought their number grows much faster. We expect to develop the transfer matrix method for the 3D fcc lattice, a natural extension from the 2D triangular lattice. We also expect to develop other applications for the method in addition to the above-mentioned application.

One area that has been studied using lattice models is designability. Some protein structures are more robust to mutations and have far more protein sequences folding to them than other protein structures. Such robust structures are said to be designable because they are more likely to be "designed" by natural selection in real organisms. Previous studies have found that, like real protein structures, lattice conformations that are selected for designability tend to have symmetries of shape and greater flexibility [6-22].

In order to find the designable conformations we thread a protein sequence onto all lattice conformations within a given shape and use a simple energy function to find the energy of each threading. If one conformation has lower energy than all other conformations we say the sequence folds to that conformation. We use binary protein sequences, made up of simple polar (P) and hydrophobic (H) residues, in order to be able to enumerate completely all sequences.

Previous studies that worked within the described framework have tested all conformations for a given lattice shape in order to find the most designable compact conformations. Here we extend that idea by generating all conformations within a group of lattice shapes. In order to make approximate comparisons we only compare shapes that have the same total number of residues (24) and also the same total number of edges (37 or 38) between vertices. In our model an edge represents either a peptide bond or a contact between two adjacent residues.

We found while comparing designability across lattice shapes that there is a marked difference in the number of sequences folding to each shape class, even after normalizing by the total number of conformations contained within each shape class. In an attempt to elucidate which features of the shape classes that could account for this difference, we compare the number of outer corners in a shape class against the total number of sequences folding to a given shape class and the radius of gyration against total number sequences folding to a given shape class.

In the case of the total number of outer corners we find a strong positive correlation with the number of sequences folding to a given shape class and in the case of radius of gyration we find a strong negative correlation with the number of sequences folding to a given shape class. This suggests that the designability signal contains information on a protein shape. Previous studies have shown that designable proteins are compact but not maximally compact [24]. We speculate that the robustness required for proteins in their natural environment perhaps could manifest itself in this deviation from compactness.

In our final chapter we employ tools from the field of computer science – machine learning algorithms [23] – in order to classify highly- and poorly-designable protein sequences. We use the 2D triangular lattice and a binary amino acid alphabet in order to model our protein structures and sequences. We enumerate all conformations within either a hexagonal or triangular shape. To determine designable conformations we thread binary sequences through all conformations within a given shape. A simple energy function allows the rapid testing of all sequences against all conformations.

Once this was completed we then have a set of binary sequences, representing real amino acid sequences. We can distinguish two subsets of sequences: those that had been designated as folding to conformations of higher and poorer designability. It was these sets of sequences that we were able to classify, often effectively with high accuracy, into groups folding into highly-designable conformations and poorly-designable conformations. The accuracy depends on how the binary sequence is represented and also depends on which machine learning algorithm is used.

To our knowledge this is the first time that machine learning algorithms have been used in the context of the designability of protein sequences and suggests that a designability signal exists and can be discerned in real protein sequences. Previous studies have already shown that some real protein structures are more designable than others, having more sequences folding to them. We are excited about the prospect of applying our study to real protein sequences and structures.

References

1. Kloczkowski A, Jernigan RL. Computer generation and enumeration of compact self-avoiding walks within simple geometries on lattices. *Computational & Theoretical Polymer Science* 1997;7:p 163-173.

2. Kloczkowski A, Jernigan RL. Efficient method to count and generate compact protein lattice conformations. *Macromolecules* 1997;30:p 6691-6694.

3. Kloczkowski A, Jernigan RL. Transfer matrix method for enumeration and generation of compact self-avoiding walks. II. Cubic lattice. *Journal of Chemical Physics* 1998;109:p 5147-5159.

4. Kloczkowski A, Jernigan RL. Transfer matrix method for enumeration ansi generation of compact self-avoiding walks. 1. Square lattices. *Journal of Chemical Physics* 1998;109:p 5134-5146.

5. A. Kloczkowski, T. Z. Sen, R. L. Jernigan, The transfer matrix method for lattice proteins-an application with cooperative interactions. *Polymer* 2004;45:p 707-716.

6. Li H, Helling R, Tang C, Wingreen N. Emergence of preferred structures in a simple model of protein folding. *Science* 1996;273:p 666-9.

7. Ejtehadi MR, Hamedani N, Seyed-Allaei H, Shahrezaei V, Yahyanejad M. Highly designable protein structures and inter-monomer interactions. *Journal of Physics A-Mathematical and General* 1998;31:p 6141-55.

8. Ejtehadi MR, Hamedani N, Seyed-Allaei H, Shahrezaei V, Yahyanejad M. Stability of preferable structures for a hydrophobic-polar model of protein folding. *Physical Review E* 1998; 57:p 3298-301.

9. Ejtehadi MR, Hamedani N, Shahrezaei V. Geometrically reduced number of protein ground state candidates. *Physical Review Letters* 1999;82:p 4723-6.

10. Shahrezaei V, Hamedani N, Ejtehadi MR. Protein ground state candidates in a simple model: An enumeration study. *Physical Review E* 1999;60:p 4629-36.

11. Shahrezaei V, Ejtehadi MR. Geometry selects highly designable structures. *Journal of Chemical Physics* 2000;113:p 6437-42.

12. Helling R, Melin R, Miller J, Wingreen N, Zeng C, Tang C. The designability of protein structures. *J Mol Graph Model* 2001;19:p 157-67.

13. Wingreen NS, Li H, Tang C. Designability and thermal stability of protein structures. *Polymer* 2004;45:p 699-705.

14. Li H, Tang C, Wingreen NS. Nature of driving force for protein folding: A result from analyzing the statistical potential. *Physical Review Letters* 1997;79:p 765-8.

15. Melin R, Li H, Wingreen NS, Tang C. Designability, thermodynamic stability, and dynamics in protein folding: A lattice model study. *Journal Of Chemical Physics* 1999;110:p 1252-62.

16. Wang TR, Miller J, Wingreen NS, Tang C, Dill KA. Symmetry and designability for lattice protein models. *Journal Of Chemical Physics* 2000;113:p 8329-36.

17. Li H, Tang C, Wingreen NS. Designability of protein structures: A lattice-model study using the Miyazawa-Jernigan matrix. *Proteins-Structure Function And Genetics* 2002;49:p 403-12.

18. Gutin AM, Abkevich VI, Shakhnovich EI. Evolution-Like Selection of Fast-Folding Model Proteins. *Proc Natl Acad Sci USA* 1995;92:p 1282-6.

19. Shakhnovich EI, Gutin AM. Engineering of Stable and Fast-Folding Sequences of Model Proteins. *Proc Natl Acad Sci USA* 1993;90:p 7195-9.

20. Shakhnovich EI. Proteins with Selected Sequences Fold Into Unique Native Conformation. *Physical Review Letters* 1994;72:p 3907-10.

21. Yue K, Dill KA. Inverse Protein Folding Problem - Designing Polymer Sequences. *Proc Natl Acad Sci USA* 1992;89:p 4163-7.

22. Cejtin C., Edler J., Gottlieb A., Helling R., Li H Fast Tree Search for Enumeration of a Lattice Model of Protein Folding *Journal of Chemical Physics* 2002; 116: p 352-359

23. Ian H. Witten and Eibe Frank "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

24. Antônio F. Pereira de Araùjo Folding protein models with a simple hydrophobic energy function: The fundamental importance of monomer inside/outside segregation. *Proc Natl Acad Sci USA*. 1999; 96(22): 12482–12487.

CHAPTER 2. COMPUTER GENERATION AND ENUMERATION OF COMPACT CONFORMATIONS WITHIN SIMPLE GEOMETRIES ON THE 2D TRIANGULAR AND 3D FCC LATTICES

A paper accepted by *Journal of Physics, Condensed Matter*

Myron Peto[b], Andrzej Kloczkowski[a], Taner Z. Sen[a,b] and Robert L. Jernigan[a,b]

Myron Peto generated most of the data and wrote up the first draft. Taner Sen generated the data for figure 16 and wrote up an explanation. Andrzej Kloczkowski and Robert Jernigan edited the paper once it was written

[a]Laurence H. Baker Center for Bioinformatics and Biological Statistics,
112 Office and Lab Bldg.
Iowa State University, Ames, IA 50011-3020

[b]Department of Biochemistry, Biophysics and Molecular Biology
Iowa State University, Ames, IA 50011-3020

Abstract

We enumerated all compact conformations within simple geometries on the 2D triangular and 3D face centered cubic (fcc) lattice. These compact conformations correspond mathematically to Hamiltonian paths and Hamiltonian circuits and are frequently used as simple models of proteins. The shapes that were studied for the 2D triangular lattice included: $m \times n$ parallelograms, regular equilateral triangles, and various hexagons. On the 3D fcc lattice we generated conformations for a limited class of skewed parallelepipeds. Symmetries of the shape were exploited to reduce the number of conformations. We compared surface to volume ratios against protein length for compact conformations on the 3D cubic lattice and for a selected set of real proteins. We also show preliminary work in

extending the transfer matrix method, previously developed by us for the 2D square and the 3D cubic lattices, to the 2D triangular lattice. The transfer matrix method offers a superior way of generating all conformations within a given geometry on a lattice by completely avoiding attrition and reducing this highly complicated geometrical problem to a simple algebraic problem of matrix multiplication.

Introduction

In spite of recent advances in computational biology, reduced models of proteins still enjoy considerable interest and applicability for studying protein structure, function, and dynamics. Globular proteins have compact structures with very tight packing of amino acids inside proteins cores due in large part to the segregation between hydrophobic and polar residues. Additionally amino acids in proteins are covalently bonded forming relatively long sequences, containing on average between few tens to few hundreds of residues. The simplest mathematical models that mimic the linear nature of the protein sequence, its tight packing in the native state and the exclusion volume effect are compact self-avoiding walks on simple lattices of finite sizes. A compact self-avoiding walk requires that each of the lattice points must be visited once and only once with no voids. Mathematically such walks are named Hamiltonian paths (an alternative nomenclature used sometimes in the literature is Hamilton paths). For regular (non-compact) self-avoiding walks some points on the lattice may be left unvisited creating voids. A compact self-avoiding walk that begins and ends at the same site is called a Hamiltonian circuit. The self-avoidance of the walks models the excluded volume condition. In lattice models of proteins each residue is usually represented by a single lattice node. Much work has been done in the past for using these models as representations of collapsed polymers and proteins.[1-11]

Native conformations of globular proteins are compact and unique. The essence of comprehending protein folding is to find, for a given sequence of amino acids, the most energetically favorable conformation. This creates extremely difficult computational problem, since the number of possible conformations grows geometrically with the length of the chain. Random search methods frequently fail to identify this single unique structure;

whereas complete enumerations, whenever feasible, are better suited and preferable for this task.

The complete enumerations of compact conformations (for both paths and circuits) within rectangles of varying sizes $n{\times}m$ on the square lattice in 2D and parallelepipeds of the size $l{\times}n{\times}m$ on the cubic lattice in 3D have been studied by us and by other authors in the past. A major obstacle in such computations for longer chains is so called 'attrition' as it becomes more and more difficult to locate unoccupied neighboring sites for the continuation the walk. To overcome this problem we developed in the past the transfer matrix method[12-16] to grow the chain not in the traditional linear way but in a piecewise way cross-section by cross-section to avoid attrition. That approach enabled us to compute all possible Hamiltonian walks and Hamiltonian circuits within rectangles of varying sizes on the square lattice and parallelepipeds on the cubic lattice.

The aim of our current work is the extension of these results to other popular lattices. The triangular lattice in 2D and the face centered cubic lattice (fcc) in 3D are especially suited for the modeling of proteins. The coordination numbers z for these lattices are 6 and 12, respectively and because of this the protein conformations generated on such lattices are more realistic than for the square (z = 4) and the cubic (z = 6) lattices. It is well known that the packing of residues inside globular proteins fits the best the fcc lattice among all other lattices[20]. Additionally the distribution of angles between vectors connecting centers of side chains of spatially neighboring residues is best fitted to 12 directional vectors in the fcc lattice[20].

For various simple geometric shapes on the 2D triangular and 3D fcc lattice we enumerate all possible compact self-avoiding walks and circuits. Figure 1 show examples of geometries studied for the triangular (Fig. 1a) and the fcc (Fig. 1b) lattice.

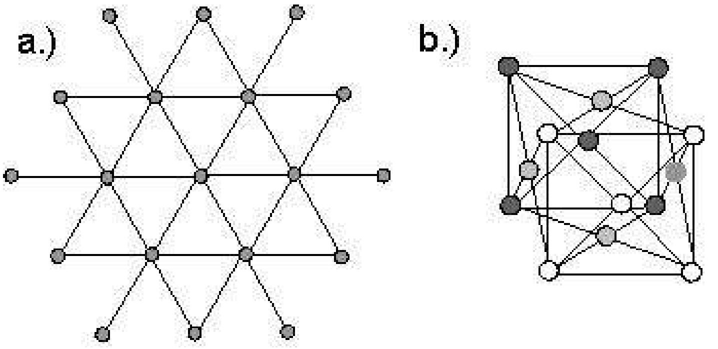


Figure 1. The 2D triangular lattice (a) and a unit cell in the face centered cubic lattice (b).

Figure 2 shows an example of a Hamiltonian circuit (Fig. 2a) and a Hamiltonian path (Fig. 2b) on the 2D triangular lattice. We enumerate all possible Hamiltonian walks and circuits within several simple geometries such as: *n*×*m* parallelograms, equilateral triangles, and several classes of hexagons of varying size. For the 3D fcc lattice we enumerate all possible walks and circuits within a limited class of skewed parallelepipeds.

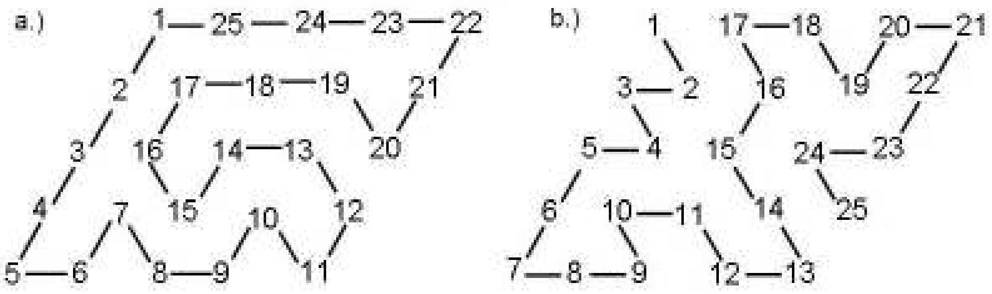


Figure 2. Examples of a Hamiltonian circuit (a) and a Hamiltonian path (b) within a parallelogram of size *5*×*5* on the 2D triangular lattice.

We take advantage of the fact that the shapes studied here exhibit symmetries. By excluding paths related by the symmetry of the shape we reduce the computer time necessary for generation. A similar approach was used earlier by us for the generation of compact conformations on the square and the cubic lattices. In the case of the 2D square and 3D cubic lattices other reductions are possible based on parity considerations. This is related to the chessboard-like nature of these lattices that can be exploited to reduce the total computational time required for the generation of compact conformations. Such a reduction isn't possible for the presently studied triangular and fcc lattices. For the square and the cubic lattice, any two nodes of a given path (that represents a polypeptide chain) that are lattice neighbors must be separated by an even number of nodes along the path. Because of this it is impossible to have a Hamiltonian circuit composed of an odd number of nodes. Indeed, one aim of our studies on the triangular and the fcc lattices is to utilize protein lattice representations that do not have such parity restrictions. In addition, the fcc lattice closely approximates the dihedral angles of real proteins[20]. The fcc lattice allows for the densest packing of hard spheres and thus the dihedral angles in densely packed proteins are associated with the fcc geometry.

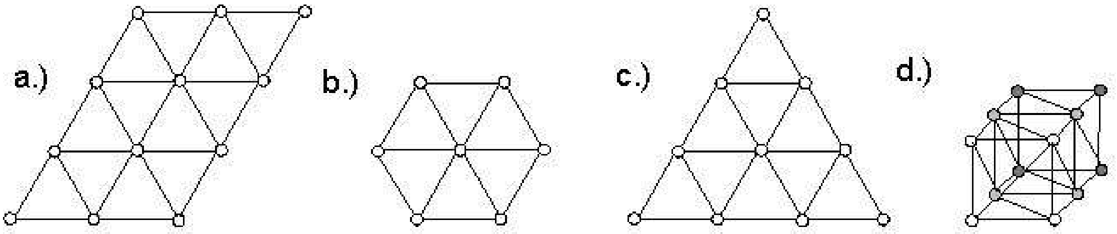


Figure 3. Examples of protein shapes on the triangular and the fcc lattices studied in the present work. (a) a *3×4* parallelogram, (b) a regular (equilateral) hexagon of side length 1 (in lattice units), (c) a regular (equilateral) triangle with sides of length 3, and (d) a *2×2×3* skewed parallelepiped on the fcc lattice.

The standard method of enumerating walks, which we employ here, uses a naïve tree-like growth algorithm. Paths are generated by adding one bond (step) in each possible way at a time and checking for possible overlaps that are not allowed. The procedure is continued until every node in the graph is visited (i.e. a Hamiltonian path is completed) or until a dead end is reached, at which point the algorithm backs up to a node where a different path (along the branches of a tree starting from that node) might be possible. This is a relatively simple to program but, especially for graphs of increasing size, suffers from the serious problem of attrition. As the number of nodes increases, fewer and fewer steps in the path generation will eventually lead to a completed path. In previous work we have developed the transfer matrix method for generating all Hamiltonian paths and Hamiltonian circuits within rectangles on the 2D square lattice and parallelepipeds on the 3D cubic lattice. Here we extend that work to the 2D triangular lattice, suggesting that it may also be possible to develop this method also for the fcc lattice and other simple lattices.

Symmetries

We exploit symmetries of shapes in order to reduce the computational cost of conformation generation. The total numbers of symmetries for all of the shapes studied in the present work (see Fig. 4 for examples) are given in *Table 1*. For the 2D triangular lattice the regular (equilateral) hexagon has the most symmetries (12); the regular (equilateral) triangle has 6 symmetries; and a parallelogram with 4 equal sides (rhomb), as well as a near equilateral hexagon with 4 sides of equal length and 2 other sides of equal length both have 4 symmetries. In three dimensions we enumerated conformations within certain classes of skewed parallelepipeds and depending on the class there can be either 2 or 4 symmetries. The use of symmetries reduces the total numbers of paths by a constant factor σ or the number of symmetries of the shape. Thus, if there are N_{total} paths without eliminating paths related by symmetry, then $N = N_{total}/\sigma$ is the number of paths after removing paths related by symmetries.

We note that, in addition to symmetries of the shape, there are also symmetries of the sequence if the graph representing the sequence is undirected. Real proteins correspond to directed graphs because of the distinction between the N-terminal and the C-terminal. It is useful however to consider undirected Hamiltonian walks on the lattice. A conformation exhibits head-tail symmetry if starting at either end of the conformation produces the same undirected graph. Fig. 5 shows an example of two conformations on the triangular lattice related by the head-tail symmetry. If the number of graphs with head-tail symmetry is N_s, then the total number of distinct directed graphs N, is related to the number of distinct undirected graphs N_u, by $N_u = (N + N_s)/2$.

Type of symmetry					
	Regular hexagon	Regular triangle	$n\times n$ parallelogram	$n\times m$ $(n \neq m)$ parallelogram	Near-regular hexagon
Identity	1	1	1	1	1
±60° rotation	2	0	0	0	0
±120° rotation	2	2	0	0	0
180° rotation	1	0	1	1	1
Reflection axial	3	0	0	0	1
Reflection diagonal	3	3	2	0	1
Total	**12**	**6**	**4**	**2**	**4**

Table 1 Symmetries for several class of shapes on the 2D triangular (above) and 3D face centered cubic (fcc) lattices (below).

Type of symmetry					
	Skewed *2x2xn* parallelepiped	Skewed *2x3xn* parallelepiped	Skewed *1tri x n* parallelepiped	Skewed *2tri x n* parallelepiped	Skewed *3tri x n* parallelepiped
Identity	1	1	1	1	1
180° rotation – facial axis	1	0	0	1	0
Reflection in an axial plane	1	0	1	1	1
Inversion	1	1	0	1	0
Total	**4**	**2**	**2**	**4**	**2**

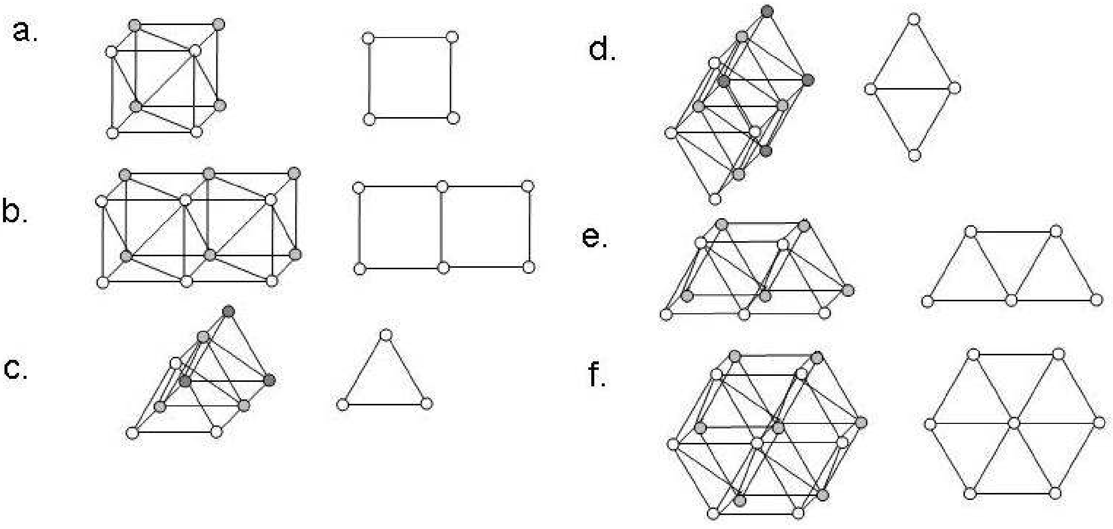


Figure 4. Shapes embedded in the 3-d fcc lattice. (a) *2x2xn*, (b) *2x3xn*, (c) *1tri x n*, (d) *2tri x n*, (e) *3tri x n,* and (f) *hex x n*. The symmetries of each are referred to in the preceding table. Cross-sections are shown on the right side.

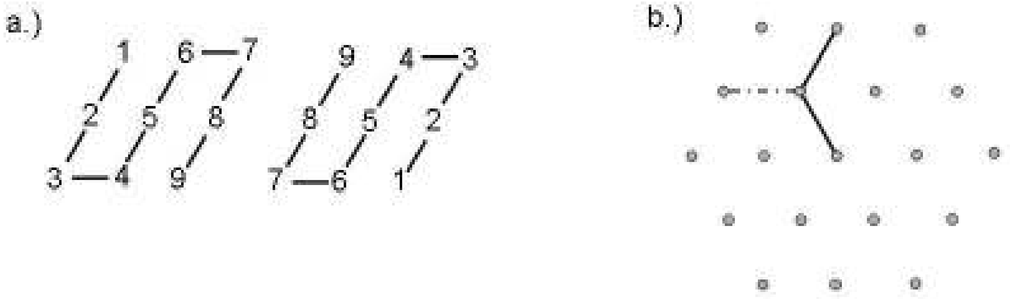


Figure 5. Dealing with symmetric conformations. (a) An example of two conformations exhibiting head-tail symmetry. The two structures are equivalent upon rotation by 180° in the plane. Shown in (b) is the method we use to eliminate symmetries. If we start our path from the central node then only one of the six equivalent nodes is chosen as the first step and only one of the two equivalent nodes as the second step (the first two steps are shown as dark lines). The other step, shown as a broken line, would produce conformations symmetrical to the first one.

We use the same method outlined in our previous work to remove symmetries. Specifically, we fix the first few steps of a path until the symmetry of the shape is broken. For example, when starting from the middle node of a regular equilateral hexagon we only need to enumerate paths with the fixed direction of the first step, since five other directions are equivalent (see Figure 5b). In addition, the second step is also fixed to break the symmetry of the shape.

Extension of the transfer matrix method to the triangular lattice

The self-avoiding walks allow the generation and enumeration of all possible compact conformations; however, due to the attrition mentioned above, the time required for these computations grows geometrically with the length of the chain. Attrition arises from the excluded volume condition together with the requirement of complete occupancy. Because of this (if the chain is grown in a traditional linear way) it becomes more and more difficult to find an unoccupied neighboring site for the subsequent step of the walk. The traditional linear chain growth method is therefore not the most efficient method for growing a chain for a compact dense system. A better approach for growing a chain is one that utilizes a piece-wise method to grow it cross-section by cross-section, using a transfer matrix method.[12-16] This method was first proposed in 1984 by Schmaltz, Hite and Klein[16] for enumerations of Hamiltonian circuits within rectangles in 2D on the square lattice. The Hamiltonian circuit (Figure 2a) is defined as a walk through all available lattice points, subject to the conditions that each site can be visited only once, and that we return in the last step back to the starting point.

The regular Hamiltonian path (Figure 2b) does not need to satisfy the second condition, and the walk (chain) has two ends. In the past work we have extended this transfer matrix method to Hamiltonian circuits in three dimensions on the cubic lattice, and to Hamiltonian paths (chains), both in two dimensions on the square lattice[15] and in three dimensions on the cubic lattice.[14] To briefly illustrate this method let us consider the enumeration of Hamiltonian circuits on a square lattice constrained to the m×n rectangular strip of width m = 4 and variable length n. Figure 5a defines all possible external connectivities to one side of the 4 points on a line. Figure 5b shows all possible distributions of bonds among the 4 points on a line, including the case with no bonds (# 1 where all bonds would be to the neighboring lines). We note that intersecting connectivities such as # 9 in Figure 5a are not allowed. Additionally connectivities #4 and #5 in Figure 5a are not allowed due to the parity reasons, so that the total number of the possible connectivity states is only six in this simple example.

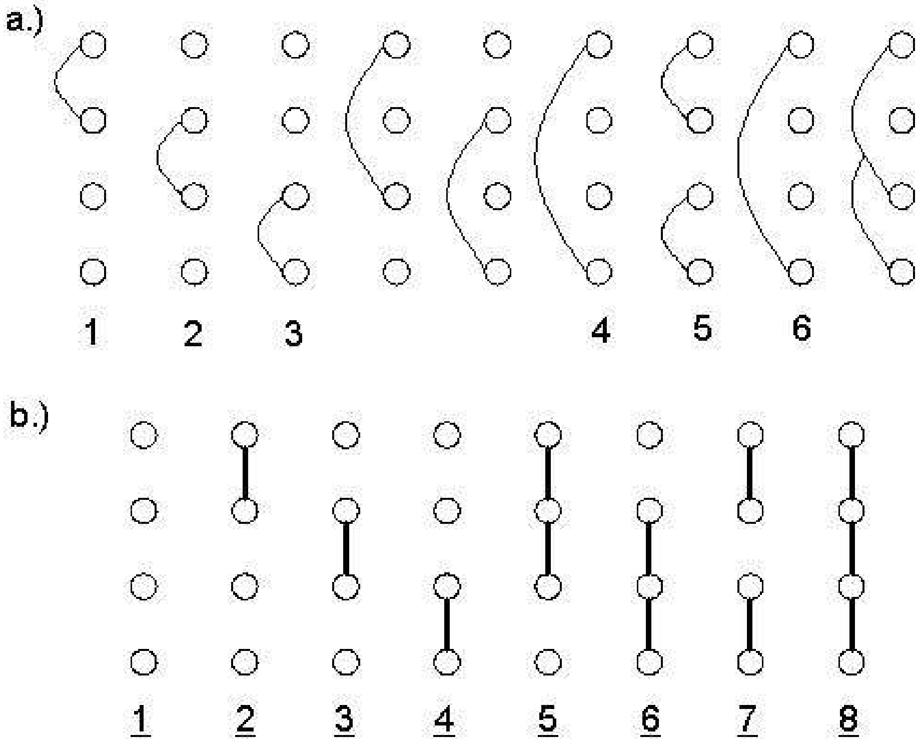


Figure 6. All possible connectivity states (a) and bond distributions (b) for generation of Hamiltonian circuits within rectangles of size 4×n.

The transfer matrix **T** is constructed by combining all connectivity states (Figure 6a) with all bond distributions (Figure 6b) and finding the resulting connectivity states formed by their combinations. The combinations, which lead to unoccupied sites, triple connections or the formation of small loops are not allowed. The element T_{ij} of the transfer matrix is zero if there is no possible transition from connectivity state i to state j. If there are possible transitions from state i to state j, then T_{ij} indicates the number of different ways to realize this transition. (For Hamiltonian circuits on the square lattice the elements T_{ij} of the matrix **T** are either 0 or 1, but in general T_{ij} can be larger than 1.) We construct the vector **u** of the starting states with elements u_i, for each connectivity state i (such as in Figure 6a) as the first state on the left in the process of building a circuit (we use a left to right convention). The number u_i identifies the number of different ways in which this may be realized. As starting states, we use the distributions of bonds (such as in Figure 6b) that do not contain any unoccupied sites

(# 7 and 8 in Figure 5b).We then determine the connectivity state to which the given distribution of vertical bonds transforms if 1) the horizontal bonds connecting to vertical bonds in the neighboring column on the right side are added and 2) a vertical cross-section passing through these newly added horizontal bonds is taken. (The distribution of bonds #7 in Figure 6b leads to the connectivity state #5 in Figure 6a, while the distribution #8 leads to the connectivity state #4.) We also construct the vector **v** of the ending states with elements v_i determining if a given connectivity state i may form a closed circuit by combining it with the distribution of vertical bonds. The exact counting of the number N_c of all possible Hamiltonian circuits on the rectangle of size m×n on the square lattice is then given by the simple formula

$$N_c = \mathbf{u}^{\mathbf{T}}(\mathbf{T})^{n-2}\,\mathbf{v}$$

with the superscript T denoting the transpose of vector **u**. If we neglect states number 4, 5 and 9 in Figure 1a and renumber the remaining states from 1 to 6 then the transfer matrix **T**, the vectors of the starting states **u** and the ending states **v** are:

$$T = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \qquad u = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \qquad v = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

In our previous work we have extended the method to Hamiltonian chains (with two ends) in two dimensions on the square lattice by generalizing the definition of the connectivity state to include the connectivities with up to 2 ends, and by generalizing bond distributions by including up to two ends.[15] We have also generalized the transfer matrix method to three dimensions (3D) on the cubic lattice both for Hamiltonian circuits and Hamiltonian paths.[14] In 2D, the cross-sections used for the generation and enumerations of

Hamiltonian paths (or circuits) were lines. In 3D the cross-sections are planes. We have written computer programs that automatically calculate the transfer matrices for paths and circuits in 2D and 3D. The only limitation is the computer memory associated with the size of the transfer matrix. The program was used to calculate transfer matrices as large as 3104×3104.

The transfer matrix method for generating and enumerating compact conformations is extremely efficient. The main advantage is that the piece-wise generation of conformations is attrition-free. Once the transfer matrix for a given cross-section is defined, the more complicated geometrical problem of conformation generation (or calculation of averages such as average energy) becomes a simple problem of matrix algebra that can be easily performed even for very long rectangles (parallelepipeds). The main difficulty of this method lies in the rapidly growing number of connectivity states for the increasing size of the cross-section, but the development of the transition matrices will be automated in order to access larger structures. Because calculations of transfer matrices are generated with a computer program, we are only limited by storage of large matrices. The algebraic formulation of the highly complicated compact self-avoiding walk problem is the principal beauty and power of this method. In the present paper we will extend the transfer matrix method to the triangular lattice.

The extension of the transfer matrix method to the triangular lattice

The triangular lattice is more difficult for studies of self-avoiding walks because its coordination number z =6 is larger than the coordination number z = 4 of the square lattice. Because of this the number of possible Hamiltonian walks and Hamiltonian circuits on the triangular lattice grows much faster with the length of the chain than for the square lattice. Additionally, for the square lattice, the parity effect associated with its chessboard-like nature (two sites that form a contact must be separated by an even number of other sites in the path) substantially reduces the number of possible connectivity states. The triangular lattice does not have this feature and all possible connectivities must be included.

We will consider Hamiltonian circuits and Hamiltonian walks within parallelograms of various sizes on the triangular lattice, such as the 5×5 parallelogram shown in Figure 2.

We will concentrate in detail on the simplest case of Hamiltonian circuits within the parallelogram of the size 3×n that will enable us to better understand the proposed transfer matrix method.

The connectivity states for the triangular lattice are defined similarly as for the square lattice by taking the cross-section along the skewed column and figuring out how various pieces of the chain are connected on the left side of this cross-section. The only generalization of this approach relative to the square lattice is that additionally to regular connectivity states similar to these shown in Figure 6a for the square lattice we need to consider situations such as that shown in Figure 7 where in the second skewed column

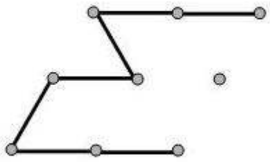


Figure 7. The extension of the transfer matrix method to the triangular lattice must take into account nodes (such as the central one in the figure) that are already occupied during the process of piece-wise building of the chain. The consideration of such nodes on the triangular lattice leads to an extension of the definition of connectivity states compared to the square lattice.

the upper and the lower sites are connected but, additionally, the site in the middle of the second column has already been occupied and therefore must be excluded in the process of choosing the transition to the connectivity states in the next cross-section. We use the symbol of an "*x*" to denote these excluded sites in the generalized connectivity states. Figure 8 shows all possible connectivity states to the 3×n parallelograms on the triangular lattice. We note that the connectivity state containing the excluded site at the top of the skewed column is not possible.

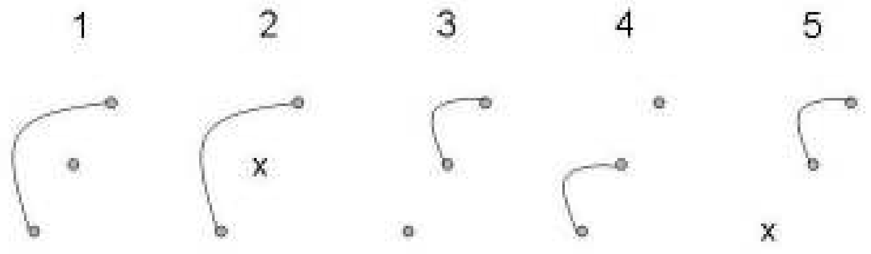


Figure 8. All possible connectivity states for Hamiltonian circuits on 3xn parallelograms on the triangular lattice. The cross symbol denotes connectivity states containing excluded nodes, such as the central node in Fig. 6.

The idea of a bond distribution, as described on cross-sections in the square lattice, must be generalized to include all bonds within a cross-section of length two. Bond distributions that would leave sites unvisited or contain short loops or triple connections are not allowed. A bond distribution here is defined in relation to the connectivity state that it corresponds to, which differs from the definition of bond distribution for the square lattice. Valid transfers are superpositions of two connectivity states that do not introduce a small cycle, a triple connection, or an orphaned site. In Figure 9 we have outlined the bond distributions as they would look in a two column format along with their corresponding connectivity states. 9b and 9c show valid and invalid transfers between two states, respectively. We have tested all connectivity states against each other to see if a valid transfer is possible.

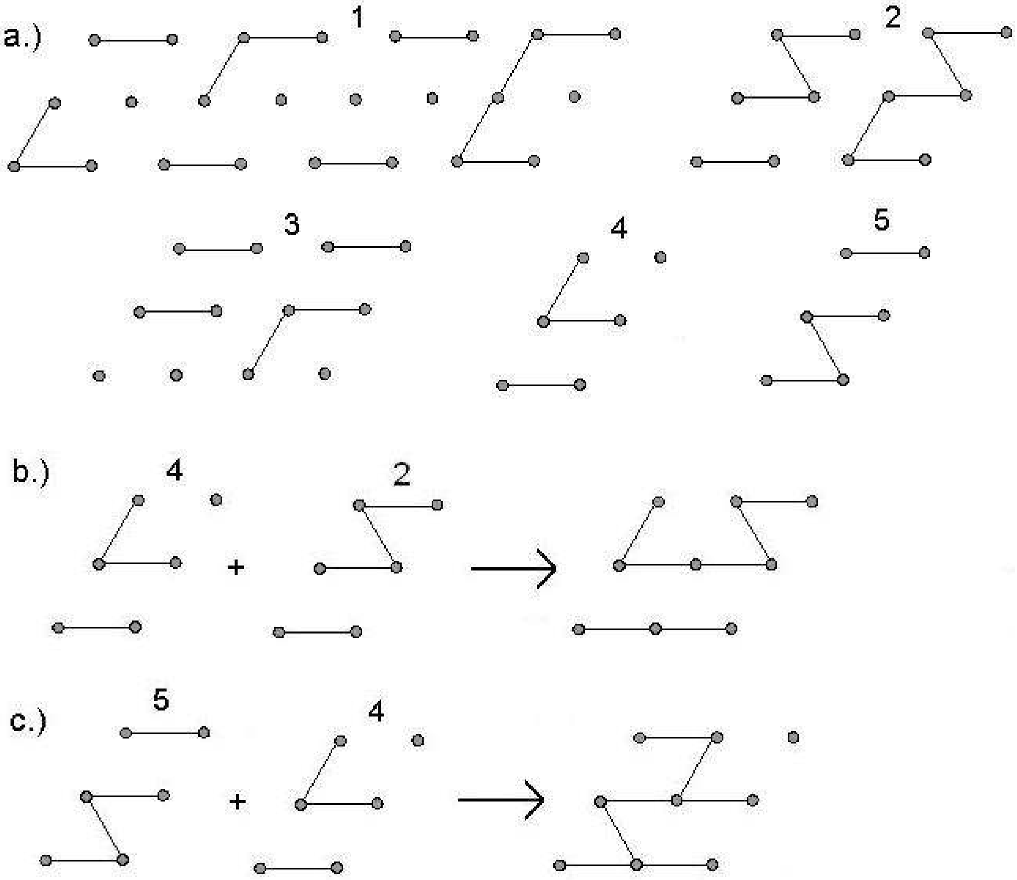


Figure 9. Bond distributions (a) for each of the 5 connectivity states. All other distributions will not lead to valid conformations (b) shows an example of a valid transfer from one state to another while (c) shows an invalid transfer from state 5 to state 4 because of a triple-bonded node.

Figure 10 illustrates all possible transitions between connectivity states for the Hamiltonian circuits on *3*×*n* parallelograms on the triangular lattice.

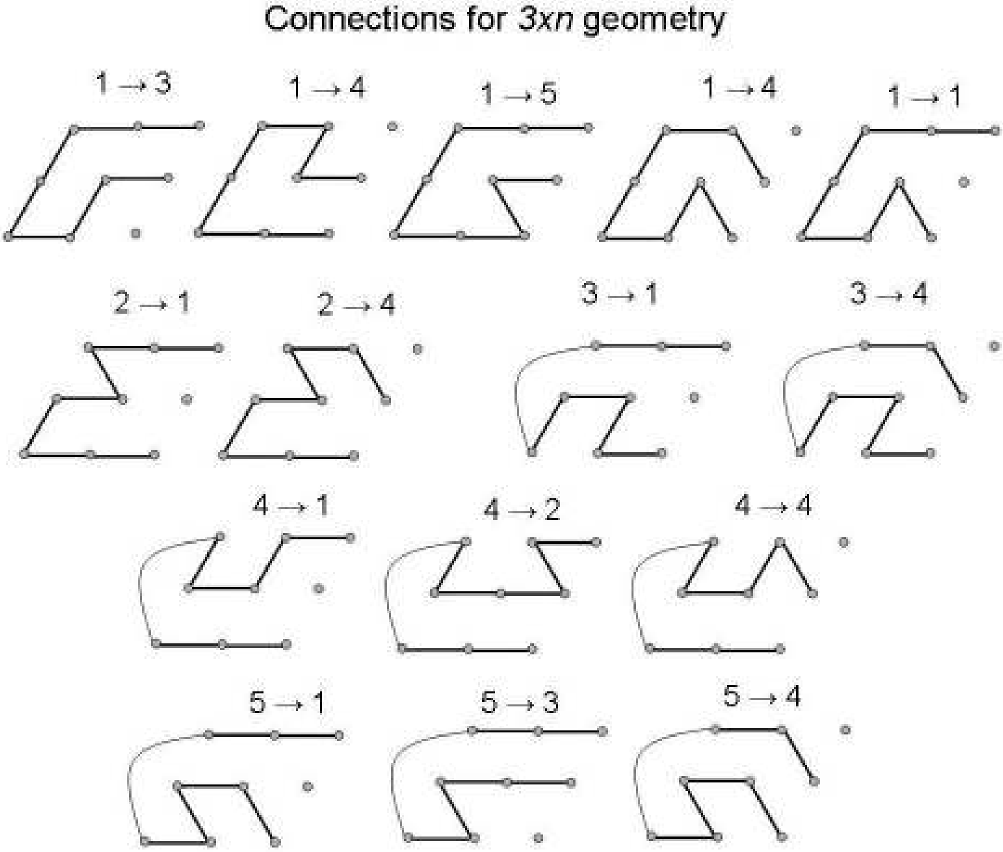


Figure 10. All possible transitions between various connectivity states for Hamiltonian circuits on *3xn* parallelograms on the triangular lattice. The notation 1→3 means the transition from connectivity state 1 (in Fig. 7) to connectivity state 3.

Because of this the transfer matrix has the following form:

$$\boldsymbol{T} = \begin{bmatrix} 1 & 0 & 1 & 2 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

The element T_{14} of the matrix has the value 2 because there are two different ways to transfer from the connectivity state 1 to the connectivity state 4 in the next cross-section by using two completely different bond distributions as shown in Figure 9.

The vector of starting states is obtained by considering all possible distributions of bonds in the first skewed column on the left and horizontal bonds joining the first column with the second one that have leave no voids in the first column and figuring out the resulting connectivity state in the second column. Figure 11a illustrates all these possibilities for Hamiltonian circuits on 3×n parallelograms on the triangular lattice.

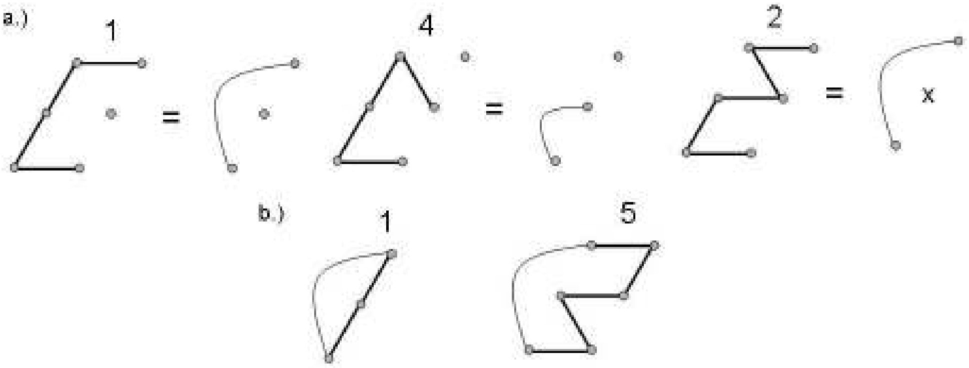


Figure 11. Connectivity states that are the starting states (a) or the ending states (b) for Hamiltonian circuits on the 3xn parallelograms on the triangular lattice.

The vector **u** of the starting states is therefore:

$$\boldsymbol{u} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

The ending connectivity states are those that lead to the closing of the circuit. Figure 11b illustrates all these ending connectivity states for 3×n parallelograms.

The vector **v** of ending states that follows from Figure 10b is

$$\boldsymbol{v} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

The number of Hamiltonian circuits for the parallelogram of length n is then obtained from the equation $N_c = \mathbf{u}^T(\mathbf{T})^{n-2}\ \mathbf{v}$. Table 2 shows the computed numbers of Hamiltonian circuits (N_n) within parallelograms of size *3*×*n* on the triangular lattice

Table 2. Conformational enumeration results

N	N_n	N_n/N_{n-1}
1	1	1.000000
2	4	4.000000
3	13	3.250000
4	44	3.384615
5	148	3.363636
6	498	3.364865
7	1676	3.365462
8	5640	3.365155
9	18980	3.365248
10	63872	3.365227
11	214944	3.365231
12	723336	3.365230
13	2434192	3.365230
14	8191616	3.365230
15	27566672	3.365230
16	92768192	3.365230
17	312186304	3.365230
18	1050578720	3.365230
19	3535439040	3.365230

The last column in Table 2 shows the ratio of the number of Hamiltonian circuits N_n/N_{n-1} for parallelograms differing in size by one column. Table 2 shows that this ratio converges rapidly with increasing size of the system.

The number of possible connectivity states for the triangular lattice grows much faster with the width of the parallelogram than for the square lattice. For example for Hamiltonian circuits on 4×n parallelograms there are 20 possible connectivity states shown in Figure 12, while for the square lattice of the same cross-section size there are only 6 states.

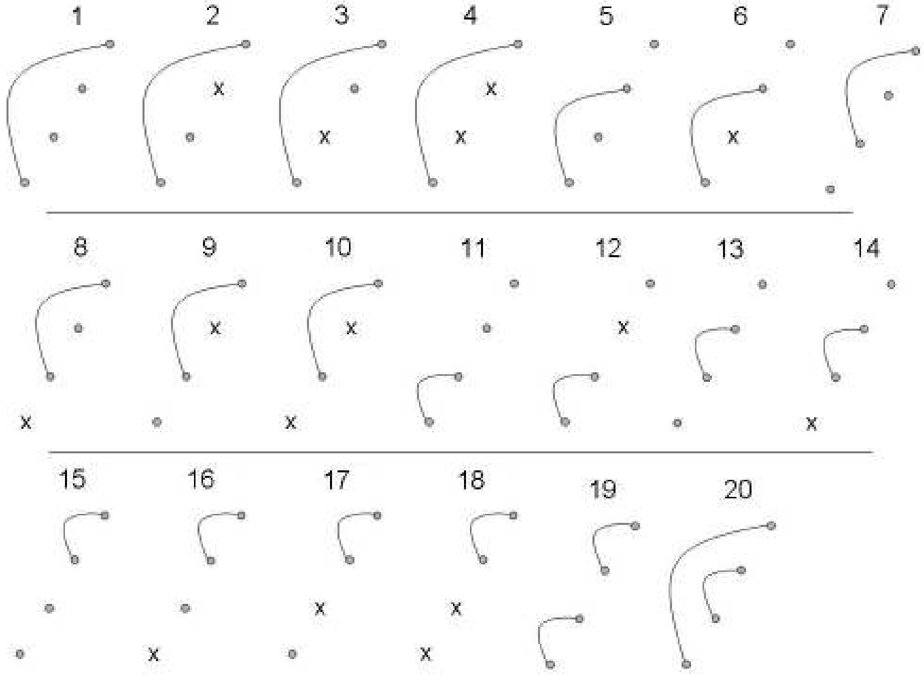


Figure 12. All possible connectivity states for Hamiltonian circuits on 4xn parallelograms on the triangular lattice.

The starting states for this case are shown in Figure 13

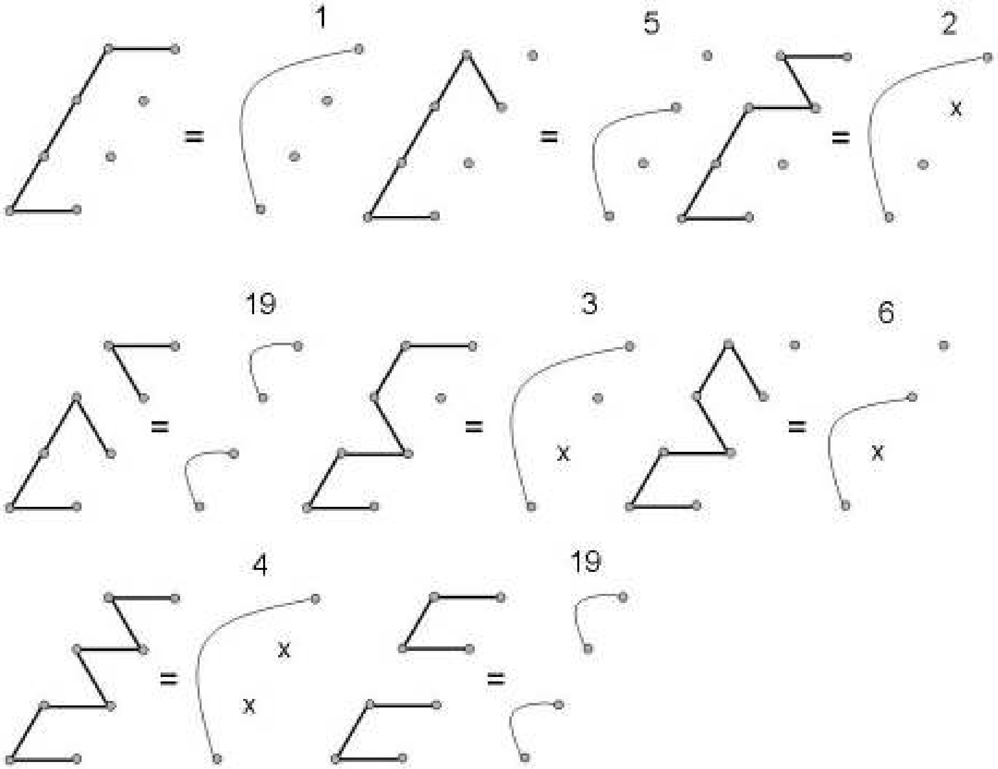


Figure 13. Connectivity states that are starting states for Hamiltonian circuits on the 4xn parallelograms on the triangular lattice.

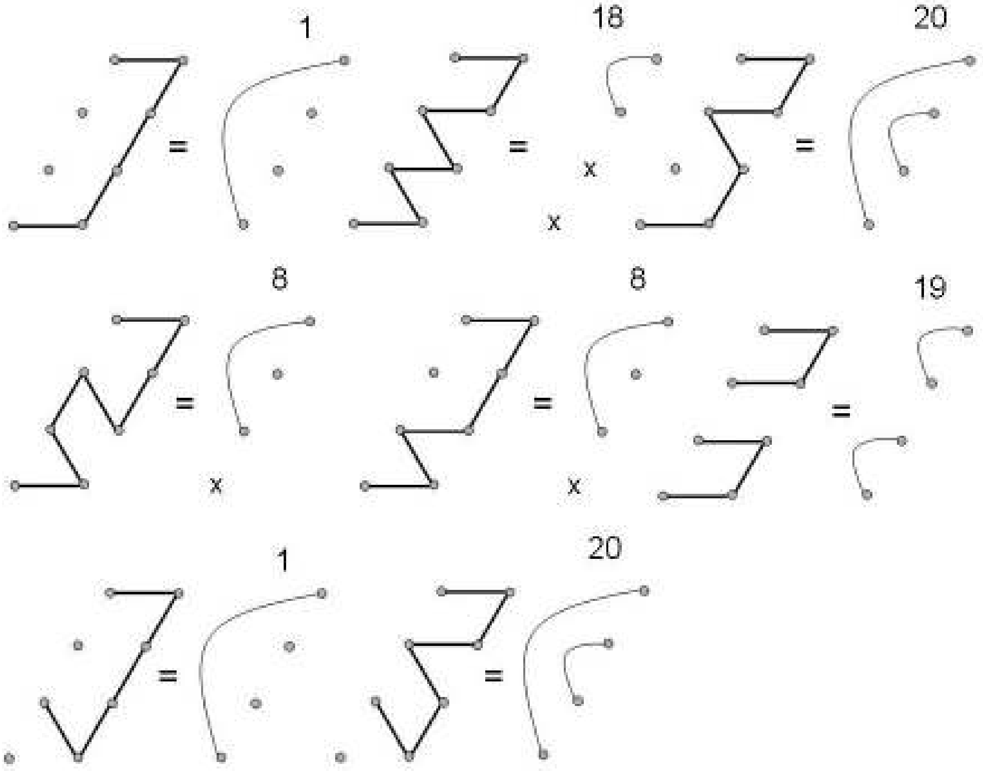


Figure 14. Connectivity states that are ending states for Hamiltonian circuits on the 4xn parallelograms on the triangular lattice.

Figure 14 shows all ending states.

We are in the process of writing the computer code that will automatically generate transfer matrices and starting and ending vectors of states for varying sizes of parallelograms for both Hamiltonian circuits and Hamiltonian paths on the triangular lattice. In the future we will generalize this method to skewed parallelepipeds on the face centered cubic lattice (fcc).

Results

We enumerated the total conformations for numerous geometries on the 2D triangular lattice and for a specific class on the fcc lattice. The totals for the various geometries in 2D and 3D are shown in Table 3. Note that our totals for the numbers of circuits in the case of $3\times n$ parallelograms matches exactly with our results taken from that of the transfer matrix method.

Table 3. Enumerations of all paths and circuits for various geometries in 2D and 3D.

(a) 2 × *n* parallelograms				(b) 3 × *n* parallelograms		
n	*N*	C_n		*n*	*N*	C_n
2	3	1		2	17	1
3	17	1		3	46	4
4	44	1		4	509	13
5	104	1		5	2525	44
6	235	1		6	11731	148
7	519	1		7	52282	498
8	1131	1		8	225105	1676
9	2448	1		9	943773	5640
10	5279	1		10	3873553	18980

(c) 4 × *n* parallelograms				(d) 5 × *n* parallelograms		
n	*N*	C_n		*n*	*N*	C_n
2	44	1		2	104	1
3	509	13		3	2525	44
4	2984	80		4	63486	549
5	63486	549		5	704218	7104
6	632663	3851		6	29534833	208200
7	6012755	26499		7	588668783	2950572
8	55267216	183521				
9	494183548	2539368				

(e) Regular triangle, *n*×*n*×*n*				(f) Hexagons, *n*×*n*×*m*		
n	*N*	C_n		*n, n, m*	*N*	C_n
2	1	1		2x2x2	10	6
3	4	1		3x3x3	20843	1284
4	38	3		2x2x3	40	40
5	656	26		2x2x4	1090	132
6	22104	474				

(g) Skewed parallelepipeds (fcc) $2\times2\times n$			(h) Skewed parallelepipeds $2\times3\times n$		
n	N	C_n	n	N	C_n
2	203	30	2	11628	381
3	8084	514	3	4301512	64758
4	296616	10136	4	1617258514	14000959
5					

(i) Skewed parallelepipeds *1 tri*$\times n$			(j) Skewed parallelepipeds *2 tri*$\times n$		
n	N	C_n	n	N	C_n
2	62	7	2	105	42
3	618	28	3	12352	726
4	5348	114	4	449942	14282
5	41836	468	5	14652475	277002
6	307764	1916	6	448917888	5380484
7	2177928	7848			
8	15020794	32144			
9	101822828	131656			

(k) Skewed parallelepipeds 3 *tri*$\times n$			(l) Skewed parallelepipeds *1 hex*$\times n$		
n	N	C_n	n	N	C_n
2	2188	103	2	137971	7588
3	173740	3722	3	183278209	4542244
4	12656898	152922			
5	818944912	6188332			

The first column in each table, n, corresponds to the length of the lattice shape. In the case of hexagons and triangles it gives an indication of the depth of the shape embedded in the lattice. The second column gives the total directed paths unrelated by symmetry. The third column gives the number of circuits. Figure 15 shows the relationship between the number of conformations and the width of the cross section *m* for various lengths of the parallelogram.

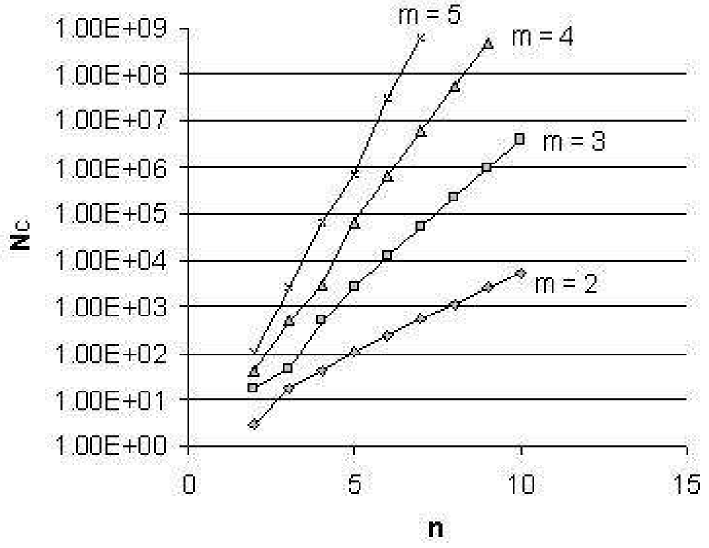


Figure 15. The plot of the number of possible Hamiltonian chains N_C vs the length *n* for varying widths *m* of the *m* X *n* parallelogram.

We also analyzed the relationship between the volume to area ratio and the total lattice sites (same as protein length) for a given lattice shape. We used the data from the earlier studies on the 3D cubic lattice as a preliminary step. We compared this with data for real protein sequences obtained from the pbd.

We used PISCES[17] to cull a data set of protein sequences with the following properties: the maximum percentage identity is less than 25%, the maximum resolution is below 2.0 Å, the maximum R-value is below 0.3, the minimum chain length is 40, and the maximum chain length is 60. We used the whole pdb[18] entry instead of separate single chains to obtain a non-redundant data set of 26 high-resolution proteins. The minimum chain length is automatically constrained to 40 by PISCES. We limited the maximum sequence length to 60, as our aim here is to analyze how the real proteins compare with a 3D square lattice, instead of observing how volume/area ratio changes with protein length. The molecular surface area and volume calculations were performed with DeepView[19].

In the 3D-square lattice representation of proteins, an increase in protein length leads to a larger number of occupied lattice nodes, and therefore a larger lattice volume. Due to geometrical constraints, with increasing protein length the volume increases at a faster rate than the area, leading to an upward slope of the volume/area ratio. This ratio can therefore be useful for assessing the reliability of any lattice model to represent protein structures. Though other descriptors, such as folding rates, and secondary structure formation, can also be used, a simple comparison of the volume/area ratio from simulations and experiments can provide a way of testing the model performance. Figure 16 shows such a comparison to validate the performance of 3D-square lattice models for representing real proteins. Here, experimental and simulation data are provided for dissimilar protein lengths. Although computational power limits the availability of results for simulated sequences up to a length of 48, the experimental data nicely integrate with the simulation data for proteins up to the length of 60 residues. The presence of noise in the experimental volume/area ratios creates a scattered plot, yet the experimental data are strongly compatible with those obtained in the simulations.

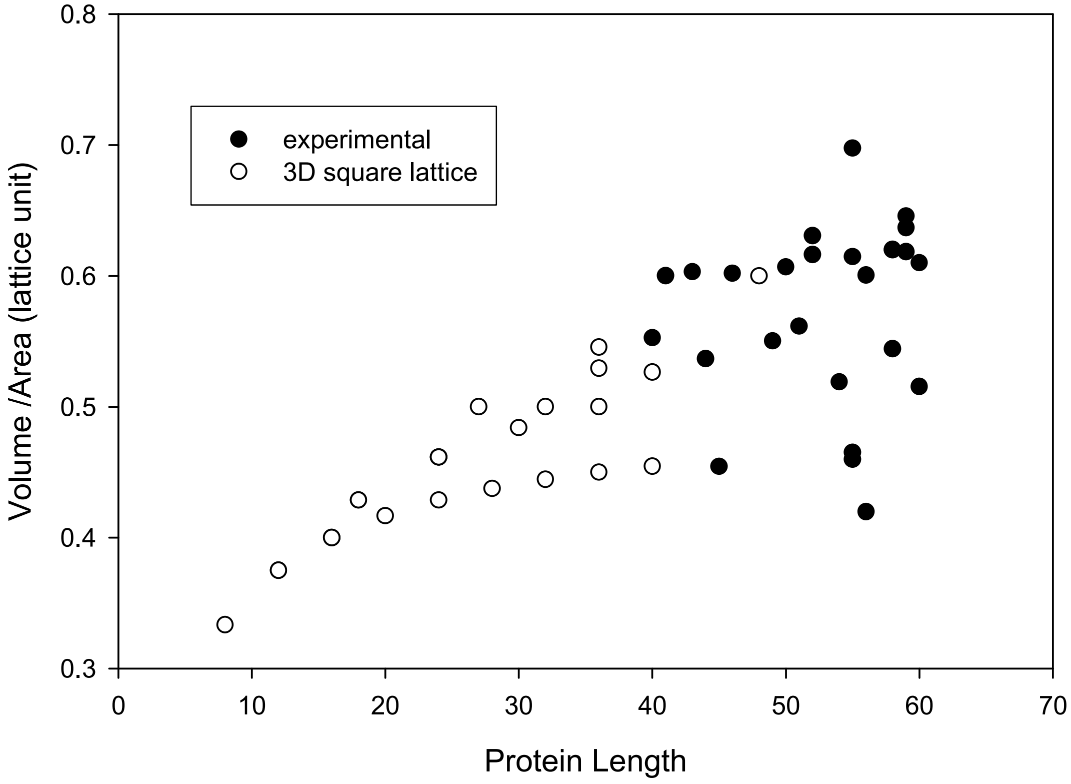


Figure 16. The volume/area ratio comparison of 3D cubic lattice conformations and real proteins as a function of protein length (length meaning number of residues).

The complementarity between experimental and simulation in Figure 16 emphasizes the utility of compact 3D-square lattice as a useful model in analyzing conformational properties of proteins constrained by geometrical requirements.

Discussion

The fact that we can enumerate the different conformations and extend the transfer matrix method lends support to the idea that we will be able to use this lattice for similar and further studies to those performed with the square and cubic lattices. As we have already developed the transfer matrix method for the triangular lattice and expect to develop it soon for the fcc lattice, proof of an algebraic method to generate conformations is not necessary.

One application for the transfer matrix method involves calculating the average energy of an ensemble of conformations[21] using differing assignments of hydrophobic and polar residues in the conformation. We intend to develop other applications in the future.

References

1. Chan HS, Dill KA. The effects of internal constraints on the configurations of chain molecules. *Journal Of Chemical Physics* 1990;92:p 3118-3135.

2. Chan HS, Dill KA. Origins of structure in globular proteins. *Proc Natl Acad Sci U S A* 1990;87:p 6388-6392.

3. Chan HS, Dill KA. Compact polymers. *Macromolecules* 2003;22:p 4559.

4. Covell DG, Jernigan RL. Conformations of folded proteins in restricted spaces. *Biochemistry* 1990;29:p 3287-3294.

5. Crippen GM. Enumeration of cubic lattice walks by contact class. *Journal Of Chemical Physics* 2000;112:p 11065-11068.

6. des Cloizeaux J, Jannink G. *Polymers in solution.* Oxford, New York: Oxford University Press; 1989.

7. Guttmann AJ, Enting IG. Solvability of some statistical mechanical systems. *Physical Review Letters* 1996;76:p 344-347.

8. Jensen I. Enumeration of compact self-avoiding walks. *Computer Physics Communications* 2003;142:p 109-113.

9. Madras N, Slade G. The self-avoiding walk. Boston: Birkhauser; 1993.

10. Shakhnovich E, Gutin A. Enumeration of all Compact Conformations of Copolymers with Random Sequence of Links. *Journal Of Chemical Physics* 1990;93:p 5967-5971.

11. Shakhnovich EI. Modeling protein folding: The beauty and power of simplicity. Folding & Design 1996;1:p R50-R54.

12. Kloczkowski A, Jernigan RL. Computer generation and enumeration of compact self-avoiding walks within simple geometries on lattices. *Computational And Theoretical Polymer Science* 1997;7:p 163-173.

13. Kloczkowski A, Jernigan RL. Efficient method to count and generate compact protein lattice conformations. *Macromolecules* 1997;30:p 6691-6694.

14. Kloczkowski A, Jernigan RL. Transfer matrix method for enumeration and generation of compact self-avoiding walks. II. Cubic lattice. *Journal Of Chemical Physics* 1998;109:p 5147-5159.

15. Kloczkowski A, Jernigan RL. Transfer matrix method for enumeration ansi generation of compact self-avoiding walks. 1. Square lattices. *Journal Of Chemical Physics* 1998;109:p 5134-5146.

16. Schmalz TG, Hite GE, Klein DJ. Compact self-avoiding circuits on two dimensional lattices. *Journal of Physics A* 1984;17:p 445-453.

17. G. Wang and R. L. Dunbrack, Jr. PISCES: a protein sequence culling server. *Bioinformatics*,S 19:1589-1591, 2003.

18. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. Nucleic Acids Research, 28 pp. 235-242 (2000).

19. N. Guex, M.C. Peitsch SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. Electrophoresis 18, 2714-2723 (1997)

20. Z. Bagci, A. Kloczkowski, R. L. Jernigan, I. Bahar The origin and extent of coarse-grained regularities in protein internal packing, *Proteins,* **53** 56-67 (2003)

21. A. Kloczkowski, T. Z. Sen, R. L. Jernigan, The transfer matrix method for lattice proteins-an application with cooperative interactions"*Polymer* **45** (2004)

CHAPTER 3. SHAPE-DEPENDENT DESIGNABILITY STUDIES OF LATTICE PROTEINS

A paper accepted by *Journal of Chemical Physics*

Myron Peto[b], Andrzej Kloczkowski[a], and Robert L. Jernigan[a,b]

Myron Peto generated all of the data and wrote up the first draft. Andrzej Kloczkowski and Robert Jernigan edited the paper once it was written.

[a]Laurence H. Baker Center for Bioinformatics and Biological Statistics,
112 Office and Lab Bldg.
Iowa State University, Ames, IA 50011-3020

[b]Department of Biochemistry, Biophysics and Molecular Biology
Iowa State University, Ames, IA 50011-3020

Abstract

One important problem in computational structural biology is protein designability, that is, why protein sequences are not random strings of amino acids but instead show regular patterns that encode protein structures. Many previous studies that have attempted to solve the problem have relied upon reduced models of proteins. In particular, the 2D square and the 3D cubic lattices together with reduced amino acid alphabet models have been examined extensively and have lead to interesting results that shed some light on evolutionary relationship among proteins. Here we perform designability studies on the 2D square lattice and explore the effects of variable overall shapes on protein designability using a binary hydrophobic-polar (HP) amino acid alphabet. Because we rely on a simple energy function that counts the total number of H-H interactions between non-sequential residues, we restrict our studies to protein shapes that have the same number of residues and also a constant number of non-bonded contacts. We have found that there is a marked difference in the

designability between various protein shapes, with some of them accounting for a significantly larger share of the total foldable sequences.

Introduction

Despite recent advances in experimental techniques and computational models for studying proteins, reduced models still enjoy considerable interest and applicability for studying fundamental features and characteristics of protein structure, function, and dynamics. Globular proteins normally have compact structures with amino acids tightly packed inside protein cores, due in large part to the segregation between hydrophobic and polar residues. Additionally, amino acids in proteins are covalently linked, forming sequences usually containing between tens to hundreds of residues. The simplest mathematical models that mimic the linear nature of the protein sequence, its tight packing in the native state and the exclusion volume effect are compact self-avoiding walks on lattices (1-18). The compact self-avoiding walk requires that each of the lattice points must be visited once and only once. Multiple visits are not allowed because of the excluded volume condition, and unvisited sites (cavities) are not allowed by the requirement of the compactness of the walk. In mathematics such walks are often called Hamiltonian paths (or Hamilton paths). A compact self-avoiding walk that begins and ends at the same site is called a Hamiltonian circuit.

The native conformations of globular proteins are compact and unique. The essence of comprehending protein folding is to find, for a given sequence of amino acids, the most energetically favorable conformation. Random search methods frequently fail to identify the single unique form; whereas complete enumerations, whenever feasible, are better suited to and preferable for this task.

In past studies of protein designability, amino acid sequences were threaded onto all possible compact conformations of a given shape and for each threading the total energy of the fold was computed based on a specified energy function (19-35). If there is a conformation that has a total energy lower than all other conformations, we assume that the sequence will fold to that specific conformation. If many different sequences fold to the same conformation we consider this conformation to be highly *designable,* and thus possibly easily unfoldable (36, 37). There are also conformations with few or even no sequences folding to them, so these have low designability, or are even completely non-designable. Additionally, many sequences do not fold uniquely; and frequently different structures can sometimes have the same lowest energy. We may however expect that such degeneracies will be reduced if a simple 2 letter (HP) amino acid alphabet were replaced by a more complex one (38). Past studies of such simple model have lead nonetheless to interesting results that shed some light on evolutionary relationship among proteins (39-42).

Previous studies that examined protein designability were mostly focused on conformations within regular lattice shapes in 2D and 3D, such as the *6×6* square or the *3×3×3* cube. Results of these studies imply the existence of few highly designable conformations among many that are less or non-designable. These results obtained for lattice proteins also suggest that, as for real proteins, designable conformations tend to exhibit symmetries. These findings show that a simple lattice model can demonstrate important traits observed for real proteins.

In an effort to further extend this model and provide greater detail regarding the structural features of protein designability, we are investigating many different shapes on the 2D square lattice. All these shapes are constrained to have both the same number of nodes (residues) and additionally the same number of non-bonded close contacts. However, lattice conformations confined by these shapes vary in their symmetries, surface characteristics, and radii of gyration. We find for a given shape differences in both the number of highly designable conformations and the total number of sequences that fold. In addition, we measure the depth of the energy well for each foldable sequence (i.e. the energy gap between the native structure and closest non-native structures) but observe only small differences in the average energy gap and average folding energy per shape class.

Methods

In an effort to extend the model to more irregular (than squares or rectangles) shapes that might more accurately mimic irregularities encountered in real proteins, we are enumerating all possible conformations within various shapes embedded in the 2D square lattice. We have performed computations for lattice proteins composed of 24 residues (nodes). The most compact shapes are the *4×6* rectangle and the *5×5* square without one of its corners (see Fig. 1A). The square lattice restricted by those shapes contains 38 edges, and because the polypeptide chain takes up 23 of these edges, this leaves 15 remaining edges for non-bonded interactions (contacts). All other shapes allow for less than 15 non-bonded contacts. In addition to studying the two most compact shapes shown in Fig. 1A, we also study various possible lattice protein shapes having 14 non-bonded contacts. This allows us to consider a larger variety of more irregular protein shapes than the two maximally compact ones. Restricting ourselves to only the most compact shapes (Fig. 1A) that allow for conformations with 15 non-bonded nearest neighbor interactions could lead to a significant oversimplification of the designability problem, and might prevent us from a more thorough examination of the relation between protein designability and shape. The shapes that allow for 14 non-bonded contacts that are studied in the present work are shown in Figure 1B. Protein shapes in Figs 1A and 1B are identified by numbers in the figure, and additionally the total numbers of different compact conformations for each shape are given there.

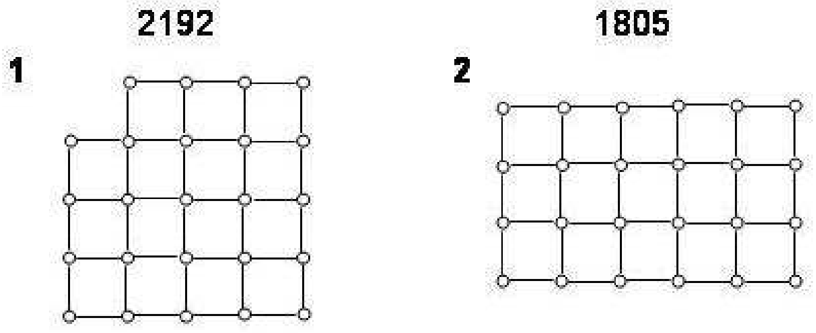


Figure 1A. The two most compact shapes comprising of 24 nodes on the square lattice, that accommodate lattice protein conformations having 15 non-bonded contacts. The shape index and the total number of all possible protein conformations for each shape are indicated.

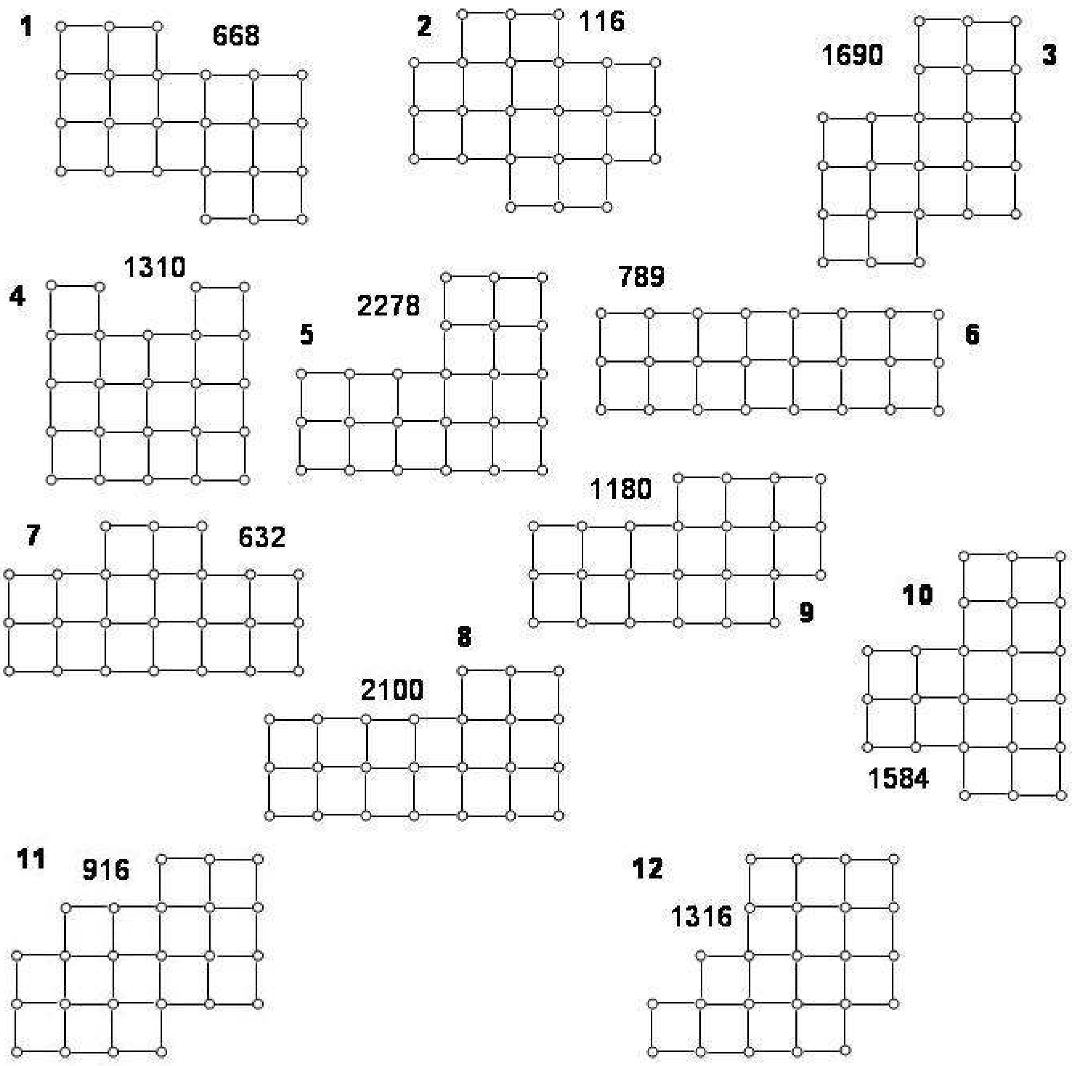


Figure 1B. Twelve different shapes composed of 24 nodes on the square nodes, which accommodate lattice protein conformations having 14 non-bonded internal contacts. The shape index and the total number of all possible protein conformations are shown for each shape.

We should note that the set of 12 shapes shown in Fig. 1B is not exhaustive, minor topological changes produce other different shapes without changing the number of vertices and edges. For example, removing two nodes (and three edges) from the upper left side of shape no. 4 and pasting them at any other possible locations on the surface produces a new shape with 24 residues and 14 non-bonded contacts. We should note, however, that there are many shapes that are excluded for parity reasons. The square lattice (and similarly the cubic lattice) has parity or even/odd characteristics, resulting from a chessboard-like structure. The allowed steps of a walk on such a lattice must connect two nodes of different parity. Because of this, the numbers of 'white' and 'black' nodes (in a chessboard terminology) must be equal for Hamiltonian circuits or may differ by zero or one for Hamiltonian walks. Shapes for which the absolute value of this difference is larger than one are not allowed. Fig. 2 shows an example of a shape that is excluded because for parity reasons; it contains 11 'white' nodes and 13 'black' nodes and therefore Hamiltonian paths (or circuits) within such a fully occupied shape are not possible.

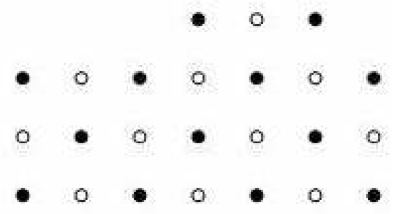


Figure 2. A shape that is impossible to fill completely with a Hamiltonian path or a circuit. Black and white nodes illustrate chessboard-like feature of the square lattice. Growing a chain will leave unoccupied nodes in all cases.

We tried to compute the number of shapes that are relatively compact by being contained within the *5×6* lattice that are the most designable. We calculate the total number of shapes with 24 residues and 14 non-bonded contacts that fit within a *5×6* rectangle on the 2D square lattice. After excluding shapes that are impossible for parity reasons and after further exclusion of shapes related by symmetry we find 92 different shapes that satisfy the *5×6* constraint. Because of limited computational resources we have not performed designability studied for all these shapes, and instead limited ourselves to sets shown in Figs 1A and 1B. Although the set of shapes in Fig. 1B is not complete, we feel that it is nonetheless adequate for the present protein designability studies and that a more complete set would add little to our findings. The set of shapes in Fig 1B contains several elongated shapes (#6, #7, #8 and #9) that do not actually fit within the *5×6* lattice; our computations have shown (see next section) that such elongated shapes are however not of high designability.

The set of shapes in Fig 1A is complete, in that there are no other shapes comprised of 24 nodes (residues) having 15 non-bonded contacts. However, such a limited number of shapes hinders a thorough investigation of the relationship between the shape and designability. The total number of all possible conformations for the two shapes in Fig. 1A is 3997.

The total number of conformations in all the different shapes in Fig. 1B is 14,579 (obtained by summing over the individual numbers of conformations for each shape). Because we study proteins with 24 residues and we are using the binary hydrophobic-polar (HP) alphabet, this amount to having 2^{24} ($\sim 3.2 \times 10^7$) different possible sequences (for chains having two distinguishable ends: C-terminal and N-terminal), each of which is threaded onto all available conformations. There are many possible energy functions even for the binary alphabet, and here we use the simplest one where each H-H non-bonded contact is given an energy score of -1.0 while all other contacts (H-P and P-P) are scored as 0. That is, E_{HH} = -1.0, E_{HP} = 0.0, E_{PP} = 0.0 in arbitrary units of energy. There is much evidence that suggests that hydrophobic interactions are the driving force in protein folding, and therefore this simple energy model captures well the essence of hydrophobic energetics in folding of real proteins.

Results

We calculate the total number of sequences that fold to each conformation with energy lower than for all other compact conformations within all shapes. Similar to previous studies, we find that there are few conformations with many sequences folding to them (i.e. highly designable conformations), and many more conformations with few or even no sequences folding to them (less designable conformations). In Fig. 3 we have shown the relationship between the number of sequences (N_s) and the logarithm ($\log_{10}$) of the number of conformations. We can see a sharp reduction in the number of conformations as the number of folding sequences increases.

In addition to this general result, we also found that certain shapes were much more accessible to designable conformations than others. The total numbers of sequences that folded to conformations confined within each shape are given in Table 1A-B. It is remarkable to observe a large diversity (differing by many orders of magnitude) in numbers of sequences folding to each shape, given that all these shapes have the same fixed numbers of vertices and edges.

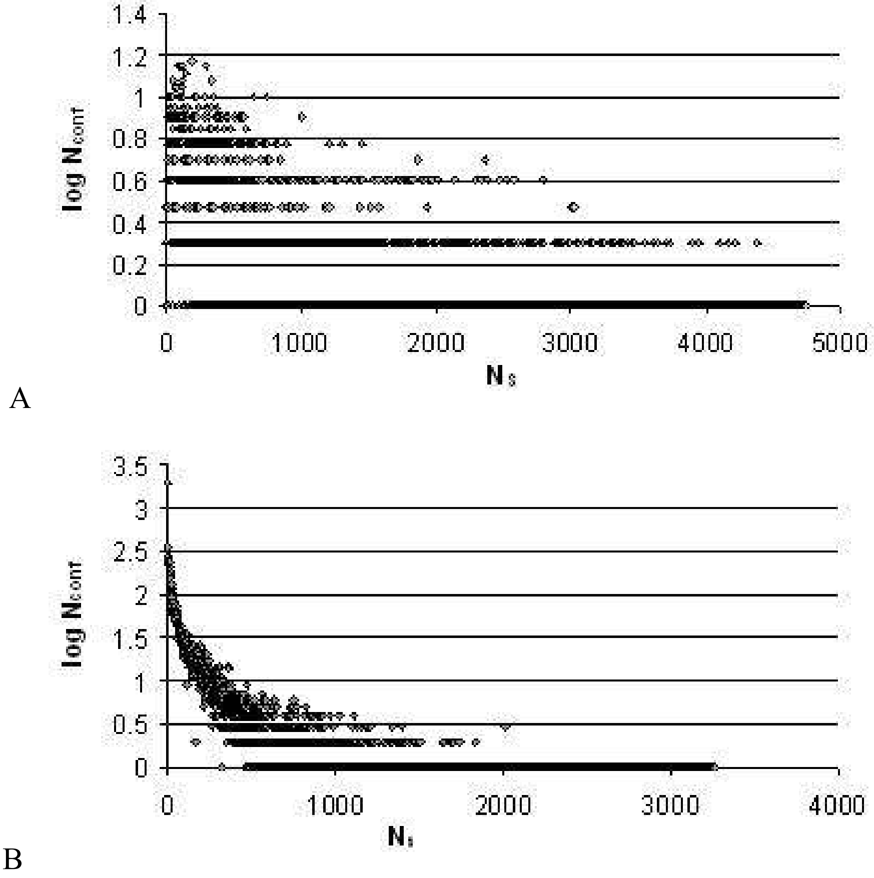


Figure 3 – The logarithm of the number of conformations plotted as the function of the total number of sequences (N_S) folding to a given conformation. (A) and (B) refer to the two different shape classes, with 15 and 14 non-bonded contacts respectively.

Such diversity could be partially explained by differences in total numbers of compact conformations for each shape. It is plausible to expect that shapes that accommodate more compact conformations might have more sequences folding to them. Because of this possibility we have normalized the number of sequences folding to a particular shape by the total number of compact conformations allowed for such shape. Such normalized numbers of sequences folding to a given shape are shown in the last column in Table 1A and B. The normalized numbers still show range from 2.0 for the shape # *6* to 760 for the shape # 1*2*. The low value (2.0) for the shape # *6* is easy to explain by its being the most elongated shape, but the unusual high designability propensity of shape # 1*2* is difficult to explain. There is a similar correlation for the two shapes with 15 non-bonded contacts, but, owing to there being only two shapes, it is difficult to draw any definitive conclusions from this evidence.

To better elucidate some of the features of the shapes that could account for the differences in designability, we have calculated the radius of gyration and the total number of corners (both inner and outer) for each shape. The mean square radius of gyration $<R_g^2>$ for each shape was computed by using the formula:

$$< R_g^{\ 2} > \ = \ \frac{1}{N^2} \sum_{i<j}^{N} (\mathbf{r}_i - \mathbf{r}_j)^2 \qquad (1)$$

where N is the number of nodes ($N = 24$) , and $\mathbf{r}_i$ is the position of the *i*-th node.

We have plotted the logarithms of the total numbers of sequences folding to particular shapes and the normalized numbers (normalized by the total number of compact conformations available for a given shape) against the mean square radius of gyration and the total number of inner and outer corners for each shape. We have studied the dependence on the total number of corners in attempting to find out how the surface characteristics of proteins influence their designability. Upon a closer examination of this problem we come to the conclusion that having corners, especially outer ones, enables energetically favorable contacts between two hydrophobic (H) residues that would not be possible for shapes without such corners. The results are shown in Figs. 4 and 5. Figure 4A shows the dependence between the mean square radius of gyration of a given shape and the logarithm of the total number of sequences folding to that shape. Fig. 4B shows a similar plot for total numbers of sequences normalized by the total number of compact conformations for each shape. It can be easily seen from these graphs that there is a strong correlation between the radius of gyration of a given shape and the logarithm of the total number of sequences folding to a particular shape. This correlation is stronger in Fig. 4B when the numbers of sequences folding to a given shape are normalized by the total number of compact conformations available for that shape. Fig. 5 show a similar plot of the total number of corners (both inner and outer ones) for each shape. There is a strong correlation between the total number of corners for a given shape and the total number of sequences folding to that shape (not shown). Similarly as in the case with the radius of gyration, the correlation increases when we normalize the total number of sequences folding to a given shape by the total number of compact conformations for that shape.

Table 2 Total and normalized numbers of sequences folding to a specific shape, corresponding to shapes with 14 non-bonded contacts.

Shape Class	Number of sequences folding to each shape class	Normalized number of conformations
1	88894	133.1
2	58495	504.3
3	201636	119.3
4	166541	127.1
5	55176	24.2
6	1563	2.0
7	99712	157.8
8	37686	17.9
9	166657	141.2
10	238385	150.5
11	381639	416.6
12	1000177	760.0

Table 3 Total and normalized numbers of sequences folding to a specific shape, corresponding to shapes with 15 non-bonded contacts.

Shape Class	Number of sequences folding to each shape class	Normalized number of conformations
1	2438869	1112.6
2	536184	297.1

We have thoroughly examined the most designable conformations and, similar to previous studies, we detect symmetries and regular secondary structure elements associated with structures of high designability. The most designable conformations for both sets of experiments are shown in Figure 6. There are 3269 and 4752 different sequences that fold to these most designable structures. The conformation A in Fig. 6 belongs to the shape # 1*2*, which is not unexpected since this shape has the highest normalized number of sequences folding to it and hence the highest density of designable conformations. The conformation B in Fig. 6 belongs to shape #1 in Fig. 1A, which is similarly densely populated with designable conformations. It is interesting that the most designable structures reveal pronounced secondary structure characteristics. It is however difficult to discern whether it is a valid representation of structural features of real proteins or an artifact resulting from the 2D square lattice representation of proteins.

We also tried to correlate shape classes with the energy difference between the conformations with the lowest energy and next lowest energy conformation. However because of the simple energy model used in our computations, for the vast majority of cases there was an energy difference of one (in arbitrary units of energy), *i.e.* the minimal possible separation between the two energy states. We have examined the average total energy, which equals the total number of H-H contacts, for all designable conformations for each shape and found only very small variations among different shapes (data not shown).

Discussion

We have generated all possible compact conformations for a variety of shapes embedded in the 2D square lattice and have performed systematic designability studies of all these conformations. We found that the different shapes vary markedly from one another in their designability propensity, with the total number of sequences folding to these shapes ranging from ~1500 to over 1,000,000. These significant differences persist even if we normalize numbers of folding sequences by the total number of possible compact conformations for each shape. We have tried to find features of the shapes that could account for this considerable difference, and have found a correlation between the mean square radius of gyration of the shape and the total number of different HP sequences folding to a given shape. This correlation is somewhat stronger after the normalization of the total number of sequences folding to a given shape by the total number of possible compact conformations within this shape. The correlation with the surface characteristics of the shapes measured by the total number of outer and inner corners is also strong, even in the case where we use total number of sequences. However, this correlation may in fact be attributable to the particular chessboard-like nature of the 2D square lattice.

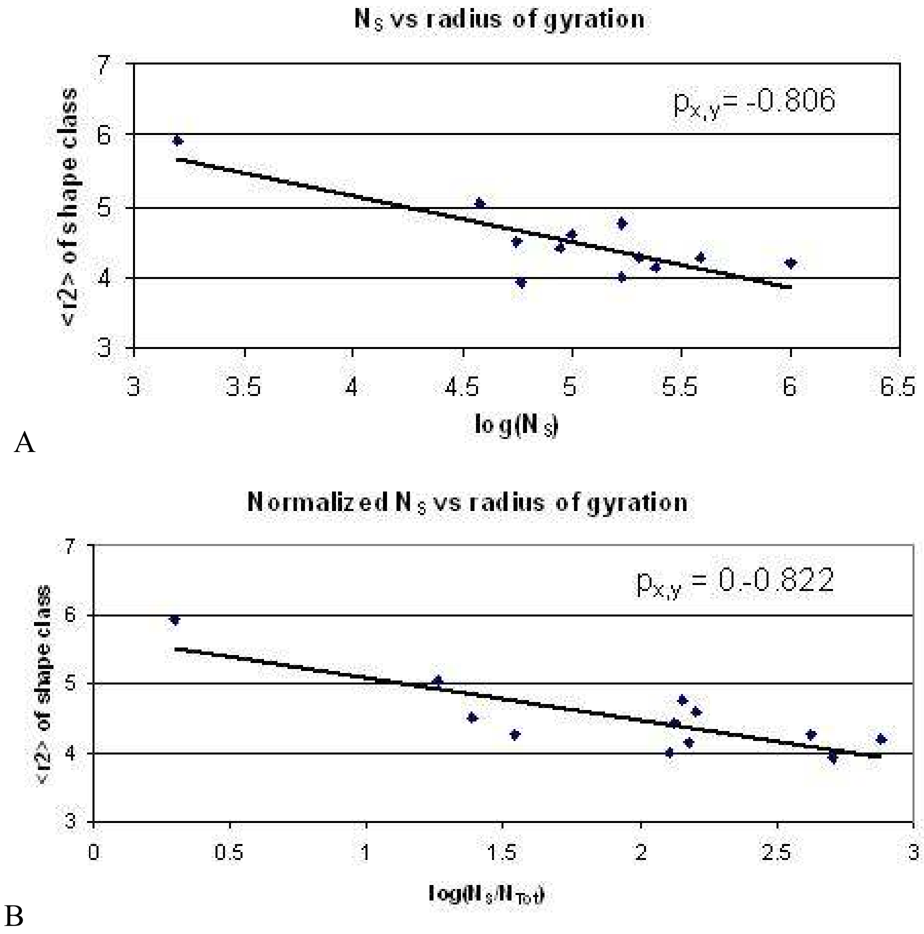
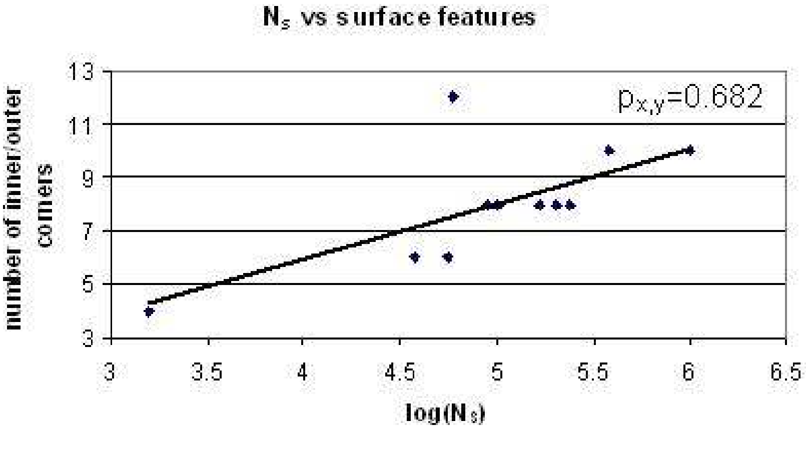


Figure 4. Correlation between the logarithm of the total number of sequences folding to a given shape and the mean square radius of gyration of this shape (A) . In the second plot (B) the number of sequences is normalized by the total number of possible compact conformations within a given shape. A linear function fits well for both plots. $p_{x,y}$ refers to the correlation coefficient, which is negative because there tend to be fewer sequences folding to shapes as the radius of gyration increases.



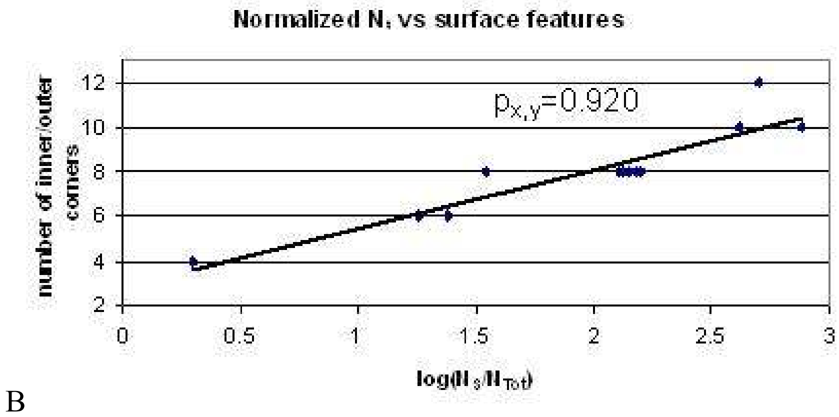


Figure 5. Correlation between the logarithm of the total number of sequences folding to a given shape (A) and the total number of inner and outer corners for this shape. (B) shows the same correlation of surface features against the total sequences normalized by the total number of possible compact conformations for a given shape.

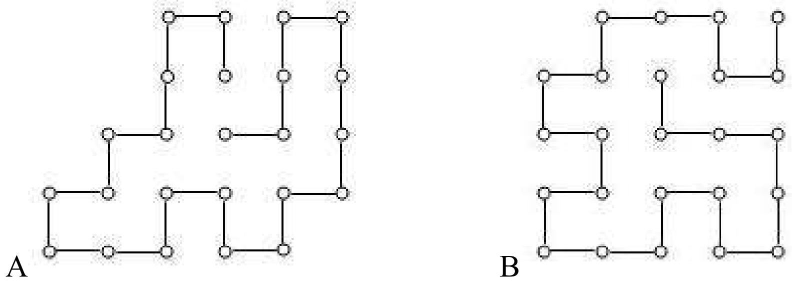


Figure 6A-B. The most designable conformation among all the shapes studied. There are 3269 different H-P sequences folding to A and 4752 sequences folding to B. A & B correspond to the shapes with 14 and 15 non-bonded contacts, respectively.

It seems possible that the differences in designability propensity between various shapes relate to the density of conformations for those shapes. Real globular proteins have dense, compact structures and we expect similar features for lattice protein models. We have explicitly tried to account for this compactness by limiting shapes that were studied to be only the most compact ones. Additionally we have compared shapes that have the same number of nodes (N = 24) and the same number of non-bonded contacts (15 contacts for two of the most compact shapes, and 14 contacts for 12 other slightly less compact shapes). A simple HP model that we use favors compact conformations in which the total number of H-H contacts are maximized, and, assuming that contacts add to the thermodynamic stability of a macromolecule, the maximization of energetically favorable H-H contacts maximizes protein stability. We may ask if there are other reasons for protein compactness. A correlation between protein designability within a given shape and the radius of gyration of this shape that we found in the present study leads us to a suggestion that perhaps proteins have evolved to minimize this value in addition to maximizing of the number of the H-H contacts. The high correlation we have found between surface features and designability may in fact suggest that proteins have evolved surfaces of optimal roughness, possibly because this lends itself to maximal compactness of the structure. However, further studies are required to rule out the possibility that our results might be artifacts of lattice used.

Similarly as in previous studies, we have found that there are relatively few highly designable conformations while the majority of compact structures generated on the square lattice are either completely non-designable or lowly designable. We have also found that most HP sequences fail to fold to a single conformation with the lowest energy. In addition, the most designable conformations tend to show some symmetry within the constraints allowed by the particular shape.

Recent studies (43, 44) have elucidated a structural determinant of protein designability for real proteins, different traces of powers of the contact matrix. These different traces correspond roughly to the average number of contacts per residue and suggest that structures with larger average number of contacts per residue are more designable. A correlation has been found between this structural determinant of designability and the size of a protein family, accounting for the evolutionary age of the family (44). It has also been discovered that proteins in thermophilic organisms, which presumably have been selected for higher thermodynamic stability, are on average more designable than those of non-thermophilic organisms (45). Our lattice protein study suggests the possibility of a correlation between protein designability and the radius of gyration (when average number of contacts per residue is used) as well as surface features in real proteins. We will attempt to examine this problem in further detail in the future work.

References

(1) Chan HS, Dill KA. The effects of internal constraints on the configurations of chain molecules. Journal Of Chemical Physics 1990;92:3118-35.

(2) Chan HS, Dill KA. Origins of structure in globular proteins. Proc Natl Acad Sci U S A 1990;87:6388-92.

(3) Chan HS, Dill KA. Compact polymers. Macromolecules 2003;22:4559.

(4) Covell DG, Jernigan RL. Conformations of folded proteins in restricted spaces. Biochemistry 1990;29(13):3287-94.

(5) Crippen GM. Enumeration of cubic lattice walks by contact class. Journal Of Chemical Physics 2000;112:11065-8.

(6) des Cloizeaux, Jannink G. Polymers in solution. Oxford, New York: Oxford University Press; 1989.

(7) Guttmann AJ, Enting IG. Solvability of some statistical mechanical systems. Physical Review Letters 1996;76:344-7.

(8) Jensen I. Enumeration of compact self-avoiding walks. Computer Physics Communications 2001;142:109-13.

(9) Kloczkowski A, Jernigan RL. Computer generation and enumeration of compact self-avoiding walks within simple geometries on lattices. Computational And Theoretical Polymer Science 1997;7(3-4):163-73.

(10) Kloczkowski A, Jernigan RL. Efficient method to count and generate compact protein lattice conformations. Macromolecules 1997;30(21):6691-4.

(11) Kloczkowski A, Jernigan RL. Transfer matrix method for enumeration ansi generation of compact self-avoiding walks. 1. Square lattices. Journal Of Chemical Physics 1998;109(12):5134-46.

(12) Kloczkowski A, Jernigan RL. Transfer matrix method for enumeration and generation of compact self-avoiding walks. II. Cubic lattice. Journal Of Chemical Physics 1998;109(12):5147-59.

(13) Madras N, Slade G. The self-avoiding walk. Boston: Birkhauser; 1993.

(14) Schmalz TG, Hite GE, Klein DJ. Compact self-avoiding circuits on two dimensional lattices. Journal of Physics A 1984;17:445-53.

(15) Shakhnovich E, Gutin A. Enumeration of all compact conformations of copolymers with random sequence of links. Journal Of Chemical Physics 1990;93(8):5967-71.

(16) Shakhnovich EI. Modeling protein folding: The beauty and power of simplicity. Folding & Design 1996;1(3):R50-R54.

(17) Mansfield ML. Monte-Carlo Studies of Polymer-Chain Dimensions in the Melt. Journal Of Chemical Physics 1982;77(3):1554-9.

(18) Mansfield ML. Unbiased sampling of lattice Hamilton path ensembles. Journal Of Chemical Physics 2006 Oct 21;125(15).

(19) Li H, Helling R, Tang C, Wingreen N. Emergence of preferred structures in a simple model of protein folding. Science 1996;273:666-9.

(20) Li H, Tang C, Wingreen NS. Are protein folds atypical? Proceedings Of The National Academy Of Sciences Of The United States Of America 1998;95(9):4987-90.

(21) Ejtehadi MR, Hamedani N, Seyed-Allaei H, Shahrezaei V, Yahyanejad M. Highly designable protein structures and inter-monomer interactions. Journal of Physics A-Mathematical and General 1998 Jul 24;31(29):6141-55.

(22) Ejtehadi MR, Hamedani N, Seyed-Allaei H, Shahrezaei V, Yahyanejad M. Stability of preferable structures for a hydrophobic-polar model of protein folding. Physical Review E 1998 Mar;57(3):3298-301.

(23) Ejtehadi MR, Hamedani N, Shahrezaei V. Geometrically reduced number of protein ground state candidates. Physical Review Letters 1999 Jun 7;82(23):4723-6.

(24) Shahrezaei V, Hamedani N, Ejtehadi MR. Protein ground state candidates in a simple model: An enumeration study. Physical Review E 1999 Oct;60(4):4629-36.

(25) Shahrezaei V, Ejtehadi MR. Geometry selects highly designable structures. Journal Of Chemical Physics 2000 Oct 15;113(15):6437-42.

(26) Irback A, Peterson C, Potthast F, Sandelin E. Design of sequences with good folding properties in coarse- grained protein models. Structure With Folding & Design 1999;7(3):347-60.

(27) Irback A, Troein C. Enumerating designing sequences in the HP model. Journal of Biological Physics 2002;28:1-15.

(28) Helling R, Melin R, Miller J, Wingreen N, Zeng C, Tang C. The designability of protein structures. J Mol Graph Model 2001;19:157-67.

(29) Wingreen NS, Li H, Tang C. Designability and thermal stability of protein structures. Polymer 2004 Jan 15;45(2):699-705.

(30) Li H, Tang C, Wingreen NS. Nature of driving force for protein folding: A result from analyzing the statistical potential. Physical Review Letters 1997;79(4):765-8.

(31) Melin R, Li H, Wingreen NS, Tang C. Designability, thermodynamic stability, and dynamics in protein folding: A lattice model study. Journal Of Chemical Physics 1999;110:1252-62.

(32) Miller J, Zeng C, Wingreen NS, Tang C. Emergence of highly designable protein-backbone conformations in an off-lattice model. Proteins-Structure Function And Genetics 2002;47:506-12.

(33) Shih C.T., Su Z.Y., Gwan J.F., Hao B.L., Hsieh C.H., Lee H.C. Mean-field HP model, designability and alpha-helices in protein structures. Physical Review Letters 2000;84:386-9.

(34) Shih C.T., Su Z.Y., Gwan J.F., Hao B.L., Hsieh C.H., Lo J.L., et al. Geometric and statistical properties of the mean-field hydrophobic-polar model, the large-small model, and real protein sequences. Physical Review E 2002;65:041923.

(35) Wang TR, Miller J, Wingreen NS, Tang C, Dill KA. Symmetry and designability for lattice protein models. Journal Of Chemical Physics 2000;113:8329-36.

(36) Dias CL, Grant M. Designable structures are easy to unfold. Physical Review E 2006 Oct;74(4).

(37) Dias CL, Grant M. Unfolding designable structures. European Physical Journal B 2006 Mar;50(1-2):265-9.

(38) Li H, Tang C, Wingreen NS. Designability of protein structures: A lattice-model study using the Miyazawa-Jernigan matrix. Proteins-Structure Function And Genetics 2002;49:403-12.

(39) Gutin AM, Abkevich VI, Shakhnovich EI. Evolution-Like Selection of Fast-Folding Model Proteins. Proceedings Of The National Academy Of Sciences Of The United States Of America 1995 Feb 28;92(5):1282-6.

(40) Shakhnovich EI, Gutin AM. Engineering of Stable and Fast-Folding Sequences of Model Proteins. Proceedings Of The National Academy Of Sciences Of The United States Of America 1993 Aug 1;90(15):7195-9.

(41) Shakhnovich EI. Proteins with Selected Sequences Fold Into Unique Native Conformation. Physical Review Letters 1994 Jun 13;72(24):3907-10.

(42) Yue K, Dill KA. Inverse Protein Folding Problem - Designing Polymer Sequences. Proceedings Of The National Academy Of Sciences Of The United States Of America 1992 May 1;89(9):4163-7.

(43) England J.L., Shakhnovich EI. Structural determinant of protein designability. Physical Review Letters 2003;90:218101.

(44) Shakhnovich BE, Deeds E, DeLisi C, Shakhnovich E. Protein structure and evolutionary history determine sequence space topology. Genome Res 2005 Mar;15(3):385-92.

(45) England J.L., Shakhnovich BE, Shakhnovich EI. Natural selection of more designable folds: a mechanism of thermophilic adaptation. Proc Natl Acad Sci U S A 2003;100:8727-31.

CHAPTER 4. USING MACHINE LEARNING ALGORITHMS TO CLASSIFY BINARY HYDROPHOBIC/POLAR PROTEIN SEQUENCES FOLDING TO HIGHLY-DESIGNABLE AND POORLY-DESIGNABLE STRUCTURES

A paper to be submitted to *BMC Bioinformatics*

Myron Peto[b], Andrzej Kloczkowski[a,b], and Robert L. Jernigan[a,b]

Myron Peto generated all of the data and wrote up the first copy. Andrzej Kloczkowski and Robert Jernigan edited the paper once it was written.

[a]Laurence H. Baker Center for Bioinformatics and Biological Statistics,
112 Office and Lab Bldg.
Iowa State University, Ames, IA 50011-3020

[b]Department of Biochemistry, Biophysics and Molecular Biology
Iowa State University, Ames, IA 50011-3020

Abstract

By using standard Support Vector Machine (SVM) with a Sequential Minimal Optimization (SMO) method of training, Naïve Bayes and other machine learning algorithms we were able to distinguish between two classes of protein sequences: those folding to highly-designable and poorly- or non-designable conformations. First, we have generated all possible compact lattice conformations for the specified shape (the hexagon or the triangle) on the 2D triangular lattice. Then we generated all possible binary hydrophobic/polar (H/P) sequences and by using a specified energy function, threaded them through all these compact conformations. If for a given sequence the lowest energy was obtained for a certain lattice conformation we assumed that this sequence folds to that conformation. Highly-designable conformations have many H/P sequences folding to them, while poorly-designable conformations have few or no H/P sequences. We classified sequences as folding to either

highly- or poorly-designable conformations. We have randomly selected subsets of the sequences belonging to high-designable and poorly-designable conformations and used them to train several different standard machine learning algorithms such as Support Vector Machine with SMO, Naïve Bayes, and Decision Tree. By using these machine learning algorithms with ten fold cross-validation we were able to classify the two classes of sequences with high accuracy - in some cases exceeding 95%.

Introduction

Elucidating the relationship between protein sequence and protein structure remains one of main unsolved problems in computational structural biology. The related specific problem is protein designability, that is, why real proteins are not random sequences of amino acids but show rather regular patterns that encode protein structures within the limited number of folds. Reduced (coarse-grained) models of proteins enjoy considerable interest and applicability for these studies. In coarse-grained models of proteins a detailed atomistic description of the structure is replaced by a much simpler view where each amino acid is represented by a single point. Additionally, theoretical models of proteins frequently replace the 20-letter amino acid alphabet with a much simpler binary hydrophobic/polar (H/P) representation and significantly restrict the conformational space by imposing lattices [1-18]. Through the use of complete enumerations of H/P sequences and compact lattice conformations it has been found that most protein sequences fold to a relatively small number of so called “highly-designable” conformations, while the remaining conformations have few, or no, sequences that fold to them [30, 38]. In the present work we use a standard H/P alphabet and a 2D triangular lattice and apply machine learning algorithms to study protein designability for such a reduced model.

Much of the past work on protein designability has focused on searching for most significant features of designable protein structures, for both lattice models and for real proteins, and relating them to energetic stability and evolution. Recently, it has been shown that proteins selected for thermal stability tend to be more highly designable, owing to their increased energetic stability [32-35]. There is also evidence suggesting that designable proteins are unfolding more easily, due to their greater flexibility [36]. Various studies have shown that designable conformations embedded on various lattices show important traits of real proteins, such as symmetrical shapes and secondary structure elements [20-31]. In addition, recent studies suggest that designable lattice structures tend to have more peptide bonds between the protein core and its surface, which increases protein flexibility [17, 36].

Those significant traits of designable conformations, found in previous works, suggested the use of machine learning algorithms to discriminate between sequences folding to highly- and poorly-designable structures. Symmetrical shapes, secondary structure elements, and extraordinary surface-core bonds can possibly show up as definitive patterns in the protein sequence; something we wanted to exploit in this study to classify sequences folding to conformations of differing designability.

In past studies of protein designability amino acid sequences were threaded onto all possible compact conformations for a given shape, and each time the total energy of the structure was computed based on a specified energy function. If, for a given amino acid sequence, there is a conformation having a total energy lower than all other conformations, it was assumed that the sequence folded to that specific structure. If many different sequences folded to the same conformation it was assumed that such a structure has high *designablility.* There were also conformations with few or even no sequences folding to them, *i.e.* with poor designability. Additionally many sequences do not fold uniquely; frequently the lowest energy is similar for different structures. We may however expect that such a degeneracy effect would rapidly diminish if a simple 2-letter (H/P) amino acid alphabet is replaced by a more complex one. Previous studies that examined the idea of protein designability were mostly focused on the conformations within regular lattice shapes in 2D and 3D, such as a 6×6 square or a $3\times3\times3$ cube. Results of these studies imply the existence of only a few highly designable conformations among a much larger number of less or non-designable structures. The results obtained for lattice proteins also suggest that, like in real proteins, designable conformations tend to exhibit structural symmetries. These findings show that a simple lattice model can demonstrate important traits that are mirrored in real proteins.

Our aim here is to extend designability studies to different shapes on the 2D triangular lattice and classify sequences folding to highly and poorly designable conformations using machine learning algorithms. The two shapes that are studied by us here are the triangle and the hexagon, shown in Figure 1. The triangular lattice with the shape of the regular hexagon in Figure 1 has 19 nodes, while the equilateral triangle contains 21 nodes. Therefore there are 2^{19} ($\cong 5.2\times 10^{5}$) and 2^{21} ($\cong 2.1\times 10^{6}$) different H/P sequences for each shape. (We have no sequence symmetry because of the difference between the C and the N terminals). Because of relatively small numbers of possible H/P sequences and the numbers of all possible compact (no voids allowed) self-avoiding walks unrelated by shape symmetries for the hexagon (20,843) and the triangle (22,104), we are able to enumerate them completely and perform complete designability computations. Similarly, as in previous studies, we find that certain distinct conformations have many sequences folding to those structures, while other have few or no sequences folding to them.

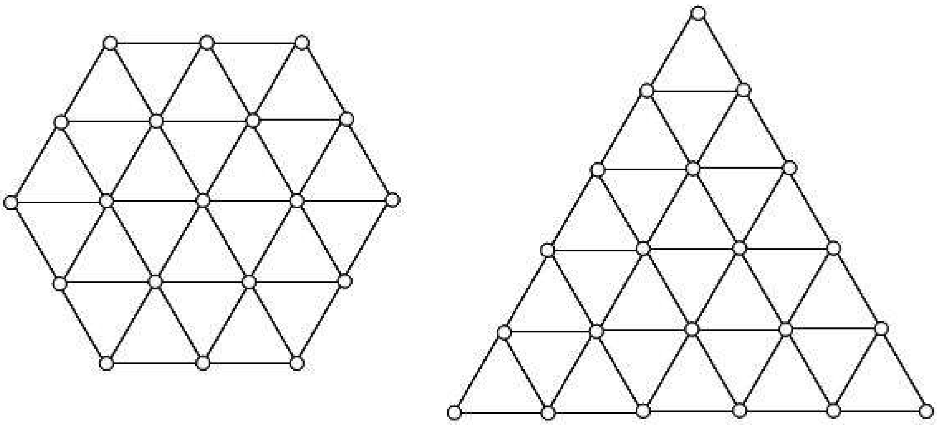


Figure 1. The hexagonal and the triangular shapes used in the designability studies in the present work. There are 20,843 different compact conformations unrelated by shape symmetries for the hexagon and 22,104 for the triangle.

After finding highly- and poorly-designable structures we then compare the sequences that folded to these two classes of conformations and tested whether we could classify them by using standard machine learning algorithms. We used the Waikato Environment for Knowledge Analysis (WEKA) software developed at http://sourceforge.net/projects/weka/ as a platform for our classification computations, testing several different algorithms such as Support Vector Machine, Naïve Bayes and Decision Tree. We first trained those statistical learning algorithms on a randomly chosen subset of our data (training set) and then checked the prediction accuracy on a remaining test set. We have performed ten-fold cross-validation experiments to eliminate possible bias. By using a Support Vector Machine with a Sequential Minimal Optimization method of training we are able to obtain highly accurate predictions, often with an accuracy exceeding 90%, depending on how the binary sequence was represented to the learning algorithm. We are quite optimistic that our approach may also be successfully applied to real proteins to distinguish protein-like sequences folding to distinct native structures from random and non-protein-like sequences that carry no significant structural signal.

Methods

The complete enumeration of all possible compact conformations for each shape was done using a backtracking algorithm generating walks on a tree that checks for all accessible nodes for the next step of the walk. If none of the nodes is available then the algorithm backtracks to the first node offering a different path. Each of nodes must be visited once and only once, voids and chain overlaps are not allowed. For longer chains this algorithm suffers from significant attrition and is less efficient than the alternative attrition-free transfer matrix approach developed by us previously [12-15]. However for the relatively short chains containing 19 or 21 nodes studied here a backtrack algorithm is much simpler to use. The energy functions that we use when calculating the total energy of the fold obtained by threading of a sequence through a conformation are based only on non-bonded nearest-neighbor contacts. Two neighbors can either be both hydrophobic (E_{HH}), one hydrophobic and one polar (E_{HP} or E_{PH}), or both polar (E_{PP}). We used a standard energy function, used in [36, 38]., that sets $E_{HH} = -2.3$, $E_{HP} = E_{PH} = -1.0$ and $E_{PP} = 0$ in energy units. This function satisfies two significant physical requirements: (i) $E_{HH} < E_{HP} < E_{PP}$ and (ii) $2E_{HP} > E_{PP} + E_{HH}$. The first requirement minimizes the number hydrophobic residues on protein surface, and the second condition allows for the separation of different amino acid types. This potential will preferentially yield overall a hydrophobic core and a polar exterior.

In order to classify the sequences folding into highly- and poorly-designable structures we used the WEKA machine learning workbench [37] and several classification algorithms, including Support Vector Machine (SVM), Decision Tree, and Naïve Bayes. As the input of the statistical learning algorithms we use two different representations of the binary amino acid sequence. Because all sequences for a given shape have the same length it was possible to simply use the binary sequence itself as input. The input vector would thus be $\boldsymbol{x} = (x_1,x_2,...,x_n)$ with elements x_i $(1\leq i \leq n)$ defined as members of the set $x \in \{0,1\}$, corresponding to either a hydrophobic or polar amino acid. In addition, we also tried using as input a percentage count of different tripeptides from the set {HHH, HHP, HPH, PHH, PPH, PHP, HPP, PPP}. The input vector is then $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)$ with x_i $(1\leq i \leq 8)$ corresponding to a percentage of the i-th tripeptide in the sequence. Encoding a sequence in this manner allows us to compare sequences of unequal length. The resulting classifiers classify a target sequence as either folding to a conformation of high designability or low designability.

The performance of our classifiers is tested using ten fold cross-validation experiments, where the data is randomly divided into ten sets, the classifier is trained on nine of the parts, and then the classifier blindly attempts to classify the remaining (known) section. The whole procedure has been repeated ten times using each of the ten sets as a test selection and the final results are compiled. The performance of a classifier can be summarized by the following metrics: *False Positives (FP)* constitute the sequences that fold to conformations of low designability but are incorrectly labeled as folding to conformations of high designability, *True Positives (TP)* are sequences that are correctly labeled as folding to conformations of high designability, *False Negatives (FN)* are sequences that are incorrectly labeled as folding to conformations of low designability, and *True Negatives (TN)* are sequences that are correctly labeled as folding to conformations of low designability.

Results

We enumerate all binary sequences and test them for possible folding to a unique native conformation with the lowest energy among all compact conformations within the given shape. As the two shapes studied by us had 19 (hexagon) and 21 (triangle) nodes this amounted to 2^{19} and 2^{21} (524,288 and 2,097,152) H/P sequences; combined with 20,843 and

22,104 conformations for each shape, repectively. We then count the number of different sequences folding to a given conformation with energy lower than all other conformations within a given shape and store the counts. These results are shown in Figure 2a for the hexagon, and Figure 2b for the triangle, where the logarithm of the number of conformations $\log N_{\mathrm{conf}}$ having N_s sequences folding to them is plotted against N_s. These two graphs express qualitatively the same ideas previously reported in earlier studies [17, 20, 21, 31, 36, 38]. There are many conformations with relatively few (or no) sequences folding to them and a rather small amount of conformations that have many sequences that fold to these structures. The later conformations are named designable conformations.

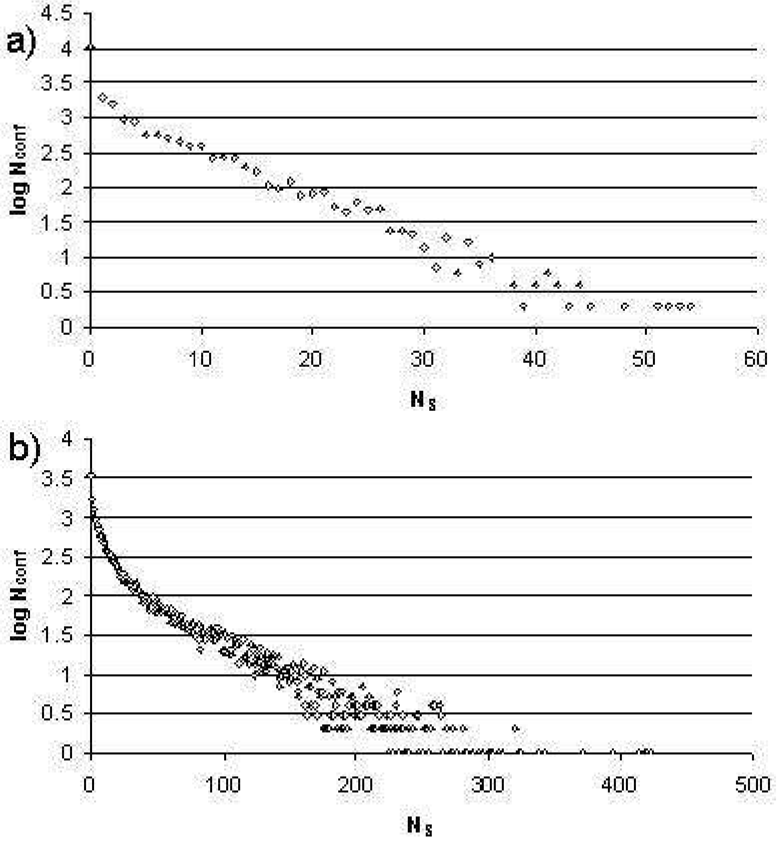


Figure 2. The logarithm of the number of conformations log N_{conf} having N_S sequences folding to them plotted as a function of N_S. a) corresponds to the data for the hexagonal shape and b) is forthe triangular shape.

Figure 3 shows the most designable conformations for both of the shapes. The most designable conformation for the hexagonal shape shows features of symmetry that have been found in . Both of the conformations contain many peptide bonds between the protein surface and the core, a feature that has been suggested to play an important role in the flexibility of proteins [36].

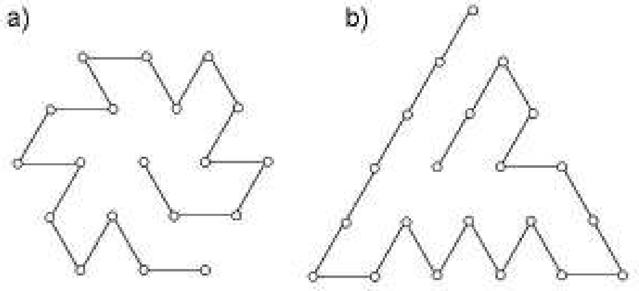


Figure 3. The most designable conformations for a) the hexagonal and b) the triangular shape. Conformation a) has 54 sequences folding to it and 11 peptide bonds connecting the protein interior with exterior; conformation b) has 423 sequences folding to it and 9 interior-exterior peptide bonds.

We have also plotted the average energy gap vs. designability of conformations for the two shapes; the results are shown in Figures 4a and 4b. The energy gap is defined as the difference between the energy of the ground state conformation and second lowest energy conformation for a given sequence. The average energy gap is the average energy gap for sequences folding to conformations of equal designability (N_S). Similarly as observed in previous studies [30, 31, 36, 38] there is a marked tendency for the energy gap to increase as we examine more designable conformations. This trend seems weaker for larger $N_{s,}$ which is probably a result of having too few conformations to obtain a reliable average. For the hexagonal shape there are less than 40 conformations more than 38 sequences folding to them, whereas there are more than 20,000 conformations with fewer sequences folding to them.

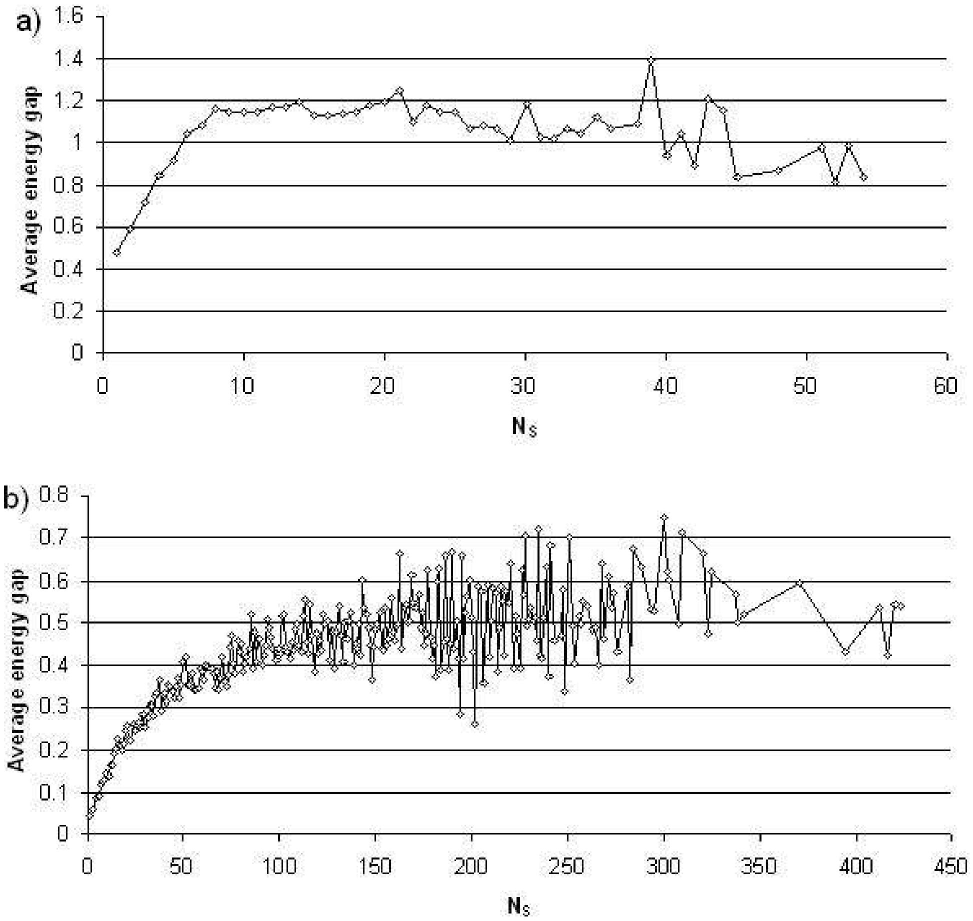


Figure 4. Average energy difference between the ground state and the next lowest energy state vs. designability for the hexagonal (a) and triangular (b) shapes. Although there is a strong visible trend towards a higher energy gap as the conformations become more designable, there are exceptions and the most designable conformations (corresponding to the largest N_s) in both cases have average energy gaps below the maximum.

It has been suggested that the number of peptide bonds connecting protein interior with exterior is related to designability, by increasing amount of protein secondary structure and allowing for easier unfolding and folding of the sequence [36]. Previous studies using lattice models have found such a relationship between the number of covalent bonds between the interior and exterior and protein designability [17, 36]. We have computed the average number of sequences folding to conformations having a specified number of peptide bonds between protein interior and exterior. The results are plotted in Figure 5 for both the hexagonal (Fig. 5a) and the triangular (Fig. 5b) shapes. Both plots show a strong dependence between the increase in the number of covalent bonds connecting protein interior with exterior and the increase in f designability, confirming earlier results of [17, 36].

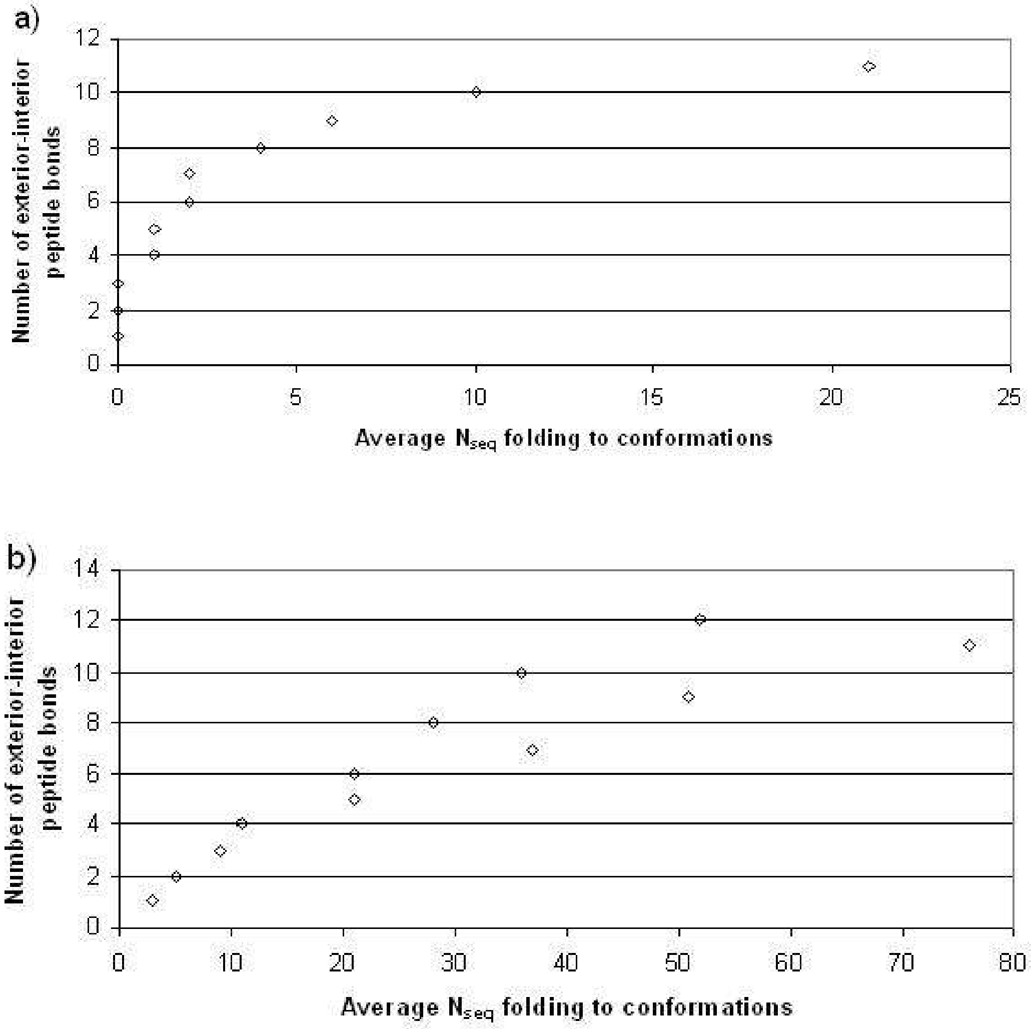


Figure 5. The average number of sequences folding to conformations having the specified number of covalent bonds connecting protein interior with exterior for a) hexagonal and b) triangular shapes.

In addition to the general results presented above, we apply machine learning algorithms to distinguish between sequences folding to highly designable and poorly designable conformations. In our first attempt we define two subsets from the set of all possible sequences: those from the bottom 10% of designable conformations and those from the top 10% of designable conformations. Because the number of sequences in both subsets differs greatly, and to reduce the computational cost, we take a random sample of sequences from each group. We could not compare sequences corresponding to different shapes, since the triangle has 21 resides while the hexagon has 19, and we performed separate computations for each shape.

Table 1 compares the accuracy of prediction obtained by using J48 Decision Tree, Naïve Bayes, and Support Vector Machine with Sequential Minimal Optimization training. As can be seen from Table 1, we are quite successful in classifying sequences based on whether they fold to highly or poorly designable conformations. All algorithms are consistently above 80% accuracy and using a Support Vector Machine results in the highest ~95% accuracy. In addition, the area under the curve (AUC), which is a measure of the overall tradeoff between the number of false positives and false negatives, was also high. This indicates that we have high accuracy with few false positives.

Table 1. Accuracy of three different machine learning prediction algorithms (J48 Decision Tree, Naïve Bayes and SVM with SMO training) using binary H/P sequences. We compare random subsets of sequences corresponding to the top 10% and the bottom 10% of designabile structures for the a) hexagon, and b) triangle. Prediction accuracy and area under the curve (AUC) for each method are shown.

	J48	Naïve Bayes	SMO
a) Sequences folding to the top 10% and the bottom 10% of designable conformations for the hexagon	95.2% correct AUC .97	86.0% correct AUC 0.96	98.2% correct AUC 0.98
b) Sequences folding to the top 10% and the bottom 10% of designable conformations for the triangle	92.7% correct AUC 0.93	82.4% correct AUC 0.92	95.0% correct AUC 0.95

We repeat the above analysis using a different representation of the binary sequence. The sequence is now represented by the percentages of all different tripeptides, which for a binary alphabet, creates 8 possibilities (HHH, HHP, HPH, PHH, HPP, PHP, PPH, and PPP). The choice of using three residue long fragments for representing the sequence seems the most natural for the triangular lattice. Using the frequency of occurrences of such short segments, gives us the advantage of being able to compare sequences of varying lengths across different shapes, allowing us to examine whether the designability traits encoded within the binary sequences are a general feature independent of the specific protein shape.

Table 2. Accuracy of three different machine learning prediction algorithms (J48 Decision Tree, Naïve Bayes and SVM with SMO training) using the frequencies of all possible short tripeptide binary segments. We compare random subsets of sequences corresponding to the top 10% and the bottom 10% of designabile structures for the a) hexagon, and b) triangle. Prediction accuracy and area under the curve (AUC) for each method are shown.

	J48	Naïve Bayes	SMO
a) Sequences folding to the top 10% and the bottom 10% of designable conformations for the hexagon	89.7% correct AUC 0.98	78.8% correct AUC 0.98	91.0% correct AUC
b) Sequences folding to the top 10% and the bottom 10% of designable conformations for the triangle	97.1% correct AUC 0.98	93.6% correct AUC 0.79	100% correct AUC 1.00

We were mildy surprised by some of the results of Table 2, namely the noticeable difference in performance between the two shapes (triangle and hexagon) as well as the 100% correct prediction accuracy using Support Vector Machine on sequences folding to conformations in the triangle shape. Neither the Naïve Bayes nor SMO algorithms give indications as to the rules that are developed and used to classify sequences as belonging to one group or another.

From the J48 decision tree results we were able to discern the tripeptide sequences that contained the most information. For the hexagon shape the two most informative tripeptides were HHH and PPP; for the triangle shape the two most informative tripeptides were PPH and HHP. This means that the percentage of HHH and PPP sequences often was used by the classifier in determining whether sequences were highly- or poorly-designable for conformations in the triangle shape (likewise PPH and HHP were used for the hexagon shape).

As to what the percentages of those particular tripeptides means, we can speculate that it could be related to the number of interior/exterior peptide bonds, as with more interior/exterior bonds we would expect more shifts between H and P, since P residues are more often found on the surface and H residues in the interior.

We are unable to formulate a reasonable justification for the differences in performances between triangle and hexagon and for the perfect classification for the triangle except to suggest that those results were an artifact of the different shapes or of the limitations of the lattice and binary alphabet.

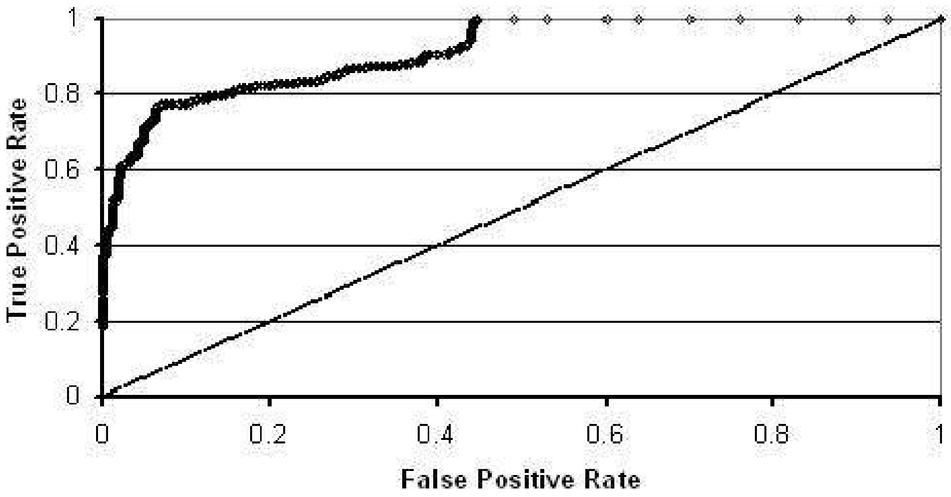


Figure 6. ROC curve for Naïve Bayes classifier. Tripeptide segments are used to classify binary sequences folding to highly- and poorly-designable conformations of the hexagonal shape. The diagonal line y=x, which we would expect if we used a classifier that randomly guessed which class to put a sequence, has been added for clarification.

In figure 6 we show a receiver operating characteristic (ROC) curve for the Naïve Bayes classifier on tripeptide sequences in the triangular shape. This plot of true sensitivity (true positives found) vs. specificity (few false positives found) gives a visual indication how our classifier performed. Qualitatively, we see that we get a large rate of true positives without having to accept many false positves. This is exactly how we want our classifier to perform and is an indication of the success of the Naïve Bayes classifier on tripeptide segments of sequences folding to conformations in the hexagonal shape.

In order to test more clearly whether the ability to distinguish between the two types of sequences is perhaps an artifact we attempt to classify highly and poorly designable sequences against random binary sequences of the same length. This means that the random sequences are of length 19 for the hexagonal shape and of length 21 for the triangular shape. As previously, we have first randomly sampled sequences from the top 10% and the bottom 10% of designable conformations for the hexagonal and triangular shapes. Then we have randomly sampled sequences of a given length (19 or 21 residues) from the set of 2^{19} (or 2^{21}) possible binary sequences and performed machine learning predictions for all these sets. Tables 3 and 4 show the results of those studies.

For each class there were approximately 300 sequences, chosen to allow a sufficient number to train the classifier but limited for the sake of computational frugality. We tested using a larger set of sequences, on the order of 1000, and saw qualitatively the same results as we see using the smaller set. The random sequences were generated using standard C++ tools. In all cases we were careful to ensure that we used two similar sized sets of sequences for our classification tests, as a disparity between the sizes of two classes can artificially improve the performance of the machine learning algorithms.

Table 3 Accuracy of machine learning predictions classifying sequences folding to the most designable conformations among random binary sequences for a) hexagonal and b) triangular shapes. Prediction accuracy and area under the curve (AUC) for each method are given.

	J48	Naïve Bayes	SMO
a) Sequences folding to the top 10% of designable structures vs. random binary sequences of length 19 for the hexagon	97.2% correct AUC 0.97	94.2% correct AUC 0.98	97.3% correct AUC 0.98

b) Sequences folding to the top 10% of designable structures vs. random binary sequences of length 21 for the triangle	90.3% correct AUC 0.91	84.4% correct AUC 0.92	95.2% correct AUC 0.95

Table 4 Accuracy of machine learning predictions classifying sequences folding to the least designable conformations among random binary sequences for a) hexagonal and b) triangular shapes. Values of prediction accuracy and area under the curve (AUC) for each method are shown.

	J48	Naïve Bayes	SMO
a) Sequences folding to the bottom 10% of designable structures vs. random binary sequences of length 19 for the hexagon	57.5% correct AUC 0.58	55.6% correct AUC 0.59	57.9% correct AUC 0.58
b) Sequences folding to the bottom 10% of designable structures vs. random binary sequences of length 21 for the triangle	50.1% correct AUC 0.50	52.3% correct AUC 0.53	56.0% correct AUC 0.56

The general result is that we are quite successful in classifying sequences that fold to highly designable structures among random sequences but are far less successful in classifying sequences folding to poorly- and non-designable structures among randomly chosen sequences. This observation is true of all machine learning algorithms and for both shapes studied .

Finally, in order to further elucidate whether binary sequences carry the shape information in their designability patterns, we attempt to classify both sequences folding to highly designable and poorly designable conformations of the hexagonal shape and the triangular shape. We have also tried machine learning methods to distinguish sequences folding to highly designable conformations folding to the hexagonal shape from poorly-designable sequences folding to the triangular shape as well as highly-designable sequences folding to the triangular shape from poorly-designable sequences folding to the hexagonal shape. Again, because we were classifying binary sequences of unequal length, we used the vector of percentages all possible tripeptides as the input to our classifiers.

Table 5 Accuracy of machine learning predictions classifying a) sequences folding to highly-designable conformations for the hexagonal and triangular shapes against sequences folding to the least designable conformations for these two shapes; b) sequences folding to the most designable conformations of the hexagonal shape against sequences folding to the least designable conformations of the triangular shape and c) sequences folding to the most designable conformations of the triangular shape against sequences folding to the least designable conformations of the hexagonal shape. Prediction accuracy and area under the curve (AUC) for each method are shown.

	J48	Naïve Bayes	SMO
a) Sequences folding to the top 10% of designable structures vs. sequences folding to the bottom 10% of designable structures for both shapes	69.5% correct AUC 0.73	65.0% correct AUC 0.69	65.6% correct AUC 0.67
b) Sequences folding to the top 10% of designable structures of hexagonal shape vs. sequences folding to the bottom 10% of designable structures in the triangular shape	98.1% correct AUC 0.99	84.9% correct AUC 0.92	87.0% correct AUC 0.87
c) Sequences folding to the top 10% of designable structures of triangular shape vs. sequences folding to the bottom 10% of designable structures in the hexagonal shape	98.0% correct AUC 0.99	65.8% correct AUC 0.70	64.3% correct AUC 0.63

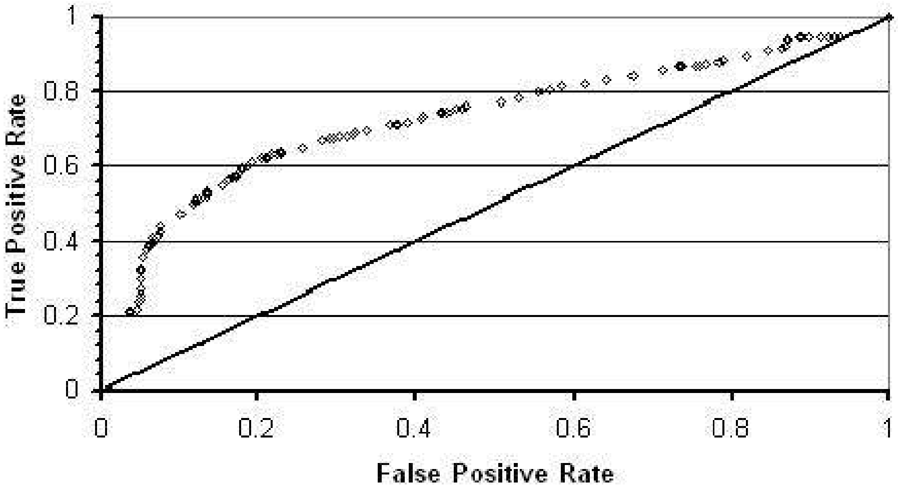


Figure 7. ROC curve for Decision Tree (J48) classifier. Tripeptide segments were used to classify binary sequences folding to highly- and poorly-designable conformations for both the hexagonal and triangular shape. The line x=y, which we would expect if we used a classifier that randomly guessed which class to put a sequence, has been added for clarification.

In figure 7 we show a receiver operating characteristic (ROC) curve for the decision tree (J48) classifier on tripeptide sequences in both the triangular and hexagonal shape. In this case our classifier performs worse than in the case of single shaped sequences (hexagonal) but is still significantly better than random guessing. This suggests there is some signal from the tripeptide segments of binary sequences folding to both shapes.

From Table 5 which shows these results we see that, although there are wide disparities among different classification algorithms and between different shapes, in general we are relatively successful in classifing sequences folding to different shapes based upon the composition of different tripeptides as the sequence representation. It is also surprising how well the Decision Tree algorithm (J48) classifies sequences folding to different shapes, in comparison to the other algorithms. When we more closely examined the tree output by the WEKA software package we found that the tri-peptide sequence PPP of three sequential polar residues carries most of the structural information. This means that the percentage of PPP tripeptide segments was a good indicator of which class (designable vs. non-designable) a sequence would fold to. Mentioned earlier, we speculate that this is related to the number of interior/exterior peptide bonds. Conformations with fewer interior/exterior bonds would have correspondingly more seqments of pure H or pure P, thus leading to the result seen.

Discussion

The protein structural designability results obtained in the present paper for two regular shapes on the 2D triangular lattice are not qualitatively different from results obtained in numerous earlier studies[17, 20, 21, 26, 31, 36, 38]. We found that designable conformations that have many sequences folding to them are relatively rare among a large number of conformations that have few or no sequences folding to them with the lowest energy. We have also found that the average energy gap between the ground state and next lowest energy state increases with increasing designability of structures; similarly as observed earlier by [30, 38].

The most interesting results obtained in our present study relate to our ability to successfully classify sequences folding to highly- and poorly-designable conformations using several standard freely available machine learning algorithms. For both studied shapes (the hexagon and the triangle) we were able to classify successfully the sequences using their full binary representation, which we may ascribe to the fact that there are relatively few highly designable conformations, and sequences folding to them probably share similar patterns in the distribution of hydrophobic and polar residues along the protein sequence.

Additionally, our further testings of sequences folding to the most designable structures among completely random sequences seems to suggest that the structural designability pattern is somehow encoded in the sequence. If the structural designability information is indeed encoded in the binary sequence we would expect to discern sequences folding to highly designable structures among random sequences much more effectively than sequences folding to poorly-designable structures. The results of our computations fully support these expectations. We could classify sequences folding to highly-designable structures among random sequences with an accuracy exceeding 90%; whereas for sequences folding to poorly- and non-designable structures our accuracy of prediction among random sequences was below 60%, i.e. not much better than random guessing (a 50% accuracy rate).

Our testing of sequences folding to designable conformations in different shapes suggests that the overall shape of the fold may also be encoded in the protein sequence.

The results presented here lend further support to the simple H/P lattice models developed for protein structural studies. Our success in classifying sequences folding to conformations in the triangular lattice, a lattice without the parity effects of the square or cubic lattice, offers evidence of the usefulness of simple models. As mentioned earlier, an interesting next step would be to test our machine learning algorithms on sequences of real proteins which fold to higher or lower designable proteins. Recent work [31-33] finds that proteins of thermophilic organisms tend also to be more designable than proteins in mesothermic organisms. We are working on classifying those two sets of protein sequences using the same tools used in this study. It would be remarkable if a designability footprint existed in real protein sequences.

References

1. Chan HS, Dill KA. The effects of internal constraints on the configurations of chain molecules. *Journal Of Chemical Physics* 1990;92:p 3118-3135.

2. Chan HS, Dill KA. Origins of structure in globular proteins. *Proc Natl Acad Sci U S A* 1990;87:p 6388-6392.

3. Chan HS, Dill KA. Compact polymers. *Macromolecules* 2003;22:p 4559.

4. Covell DG, Jernigan RL. Conformations of Folded Proteins in Restricted Spaces. *Biochemistry* 1990; 29:p 3287-3294.

5. Crippen GM. Enumeration of cubic lattice walks by contact class. *Journal Of Chemical Physics* 2000;112:p 11065-11068.

6. J.des Cloizeaux, Jannink G. *Polymers in solution.* Oxford, New York: Oxford University Press; 1989.

7. Guttmann AJ, Enting IG. Solvability of some statistical mechanical systems. *Physical Review Letters* 1996;76:p 344-347.

8. Jensen I. Enumeration of compact self-avoiding walks. *Computer Physics Communications* 2003;142:p 109-113.

9. Madras N, Slade G. The self-avoiding walk. Boston: Birkhauser; 1993.

10. Shakhnovich E, Gutin A. Enumeration of all Compact Conformations of Copolymers with Random Sequnce of Links. *Journal Of Chemical Physics* 1990;93:p 5967-5971.

11. Shakhnovich EI. Modeling protein folding: The beauty and power of simplicity. Folding & Design 1996;1:p R50-R54.

12. Kloczkowski A, Jernigan RL. Computer generation and enumeration of compact self-avoiding walks within simple geometries on lattices. *Computational And Theoretical Polymer Science* 1997;7:p 163-173.

13. Kloczkowski A, Jernigan RL. Efficient method to count and generate compact protein lattice conformations. *Macromolecules* 1997;30:p 6691-6694.

14. Kloczkowski A, Jernigan RL. Transfer matrix method for enumeration and generation of compact self-avoiding walks. II. Cubic lattice. *Journal Of Chemical Physics* 1998;109:p 5147-5159.

15. Kloczkowski A, Jernigan RL. Transfer matrix method for enumeration and generation of compact self-avoiding walks. 1. Square lattices. *Journal Of Chemical Physics* 1998;109:p 5134-5146.

16. Schmalz TG, Hite GE, Klein DJ. Compact self-avoiding circuits on two dimensional lattices. *Journal of Physics A* 1984;17:p 445-453.

17. Cejtin C., Edler J., Gottlieb A., Helling R., Li H Fast Tree Search for Enumeration of a Lattice Model of Protein Folding *Journal of Chemical Physics* 2002; **116** p352-359

18. Mansfield ML 2006 Unbiased sampling of lattice Hamiltonian path ensembles *J. Chem. Phys.* **125** (15)

19. N. Guex, M.C. Peitsch SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. Electrophoresis 18, 2714-2723 (1997)

20. Shahrezaei V, Ejtehadi MR, Geometry selects highly designable structures Journal of Chemical Physics 113 (15): 6437-6442 OCT 15 2000

21. Shahrezaei V, Hamedani N, Ejtehadi MR, Protein ground state candidates in a simple model: An enumeration study, *Physical Review E* 60 (4): 4629-4636 Part B OCT 1999

22. Ejtehadi MR, Hamedani N, Shahrezaei V Geometrically reduced number of protein ground state candidates *Physical Review Letters* 82 (23): 4723-4726 JUN 7 1999

23. Ejtehadi MR, Hamedani N, Seyed-Allaei H, et al. Highly designable protein structures and inter-monomer interactions *Journal of Physics A - Mathematical and* General 31 (29): 6141-6155 JUL 24 1998

24. Ejtehadi MR, Hamedani N, Seyed-Allaei H, et al. Stability of preferable structures for a hydrophobic-polar model of protein folding *Physical Review E* 57 (3): 3298-3301 Part B MAR 1998 re accounted for in real proteins.

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

26. Li, H Tang C, Wingreen N. S. Nature of driving force for protein folding: A result from analyzing the statistical potential *Physical Review Letters* 97 (4) 765-768 1997

27. Shakhnovich E. I. Proteins with selected sequences fold into unique native conformation *Physical Review Letters* 72 (24) 3907-3910 1994

28. Gutin A. M., Abkevich V. I., Shakhnovich E. I. Evolution-like selection of fast-folding model proteins *Proc. Natl. Acad. Sci.* 1995;92:p 1281-1286

29. Yue K., Dill K. A. Inverse protein folding problem: designing polymer sequences *Proc. Natl. Acad. Sci.* 1992;89:p 4163-4167

30. Li H., Tang C., Wingreen N. Designability of protein structures: a lattice-model study using the Miyazawa-Jernigan matrix *PROTEINS: Structure, Function and Genetics* (49) 403-412 2002

31. Wingreen N., Li H., Tang C. Designability and thermal stability of protein structures *Polymer* 45 699-705 2004

32. Shakhnovich B., Deeds E., Delisi C., Shakhnovich E. I. Protein structure and evolutionary history determine sequence space topology *Genome Research* 15: 385-392 2005

33. England J. L., Shakhnovich B., Shahknovich E. I., Natural selection of more designable folds: A mechanism for thermophilic adaptation *Proc. Natl. Acad. Sci.* 2003;100:p 8727-8731

34. Berezovsky I. N., Shahknovich E. I., Physics and evolution of thermophilic adaptation *Proc. Natl. Acad. Sci.* 2005;102:p 12742-12747

35. Berezovsky I. N., Zeldovich K. B., Shahknovich E. I., Positive and Negative Design in Stability and Thermal Adaptation of Natural Proteins *PLOS Computational Biology* 2007; 3: p498-507

36. Dias C. L., Grant M., Designable Structures Are Easy to Unfold *Phys Review E* 2006; **74**: 42902(4)

37. Ian H. Witten and Eibe Frank "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

38. Li H., Helling R., Tang C., Wingreen N., Emergence of Preferred Structures in a Simple Model of Protein Folding *Science* 1996; **273** p666-669

CHAPTER 5. CONCLUSION

The results obtained here suggest that simplified lattice/HP models of proteins are still highly useful. The ability to enumerate completely a set of conformations and a set of sequences leads to analyses and conclusions that would otherwise be difficult or impossible with more detailed or continuous models. We have expanded the existing knowledge on lattice models of proteins in three significant ways.

1. We took the transfer matrix method developed originally for the square and cubic lattices and applied it in a novel way to a new lattice with higher coordination number, the 2-D triangular lattice, that has no limitations due to parity. We have already found an application for this method in studying averages of conformational ensembles. In the future we expect to develop more applications for this method and also to extend it to the 3D fcc lattice.
2. We undertook a novel study examining the general shapes of lattice conformations and how that shape influences the designability of proteins. Somewhat unexpectedly, even after holding the total numbers of bonds and residues constant and accounting for the differences in total numbers of conformations available for a given shape, we see significant differences in the designabilities of various shape classes. We attempted to account for those differences, in the form of different radii of gyration and other physical traits. Extending this work to other lattices would probably be too computationally expensive to be feasible, as the number of shape classes would grow much faster for lattices with higher coordination numbers. However, it should be possible to compare radii of gyration and surface features against designability for real proteins.
3. We applied machine learning algorithms to the designability issue and found that we can distinguish quite well between sequences folding to highly- and poorly-designable conformations. We attempted to account for possible artifacts by using a sequential three peptide representation, comparing against random sequences, and by comparing sequences folding to different shapes (triangle and hexagon). In all cases were we successful in our ability to classify the two categories of sequences. As with

the previous study of protein shapes, an important next step in this direction would be to apply our method to real protein structures and sequences. It would be interesting if we were able to detect designability signals discernable for real amino acid sequences. If this were indeed the case, it would open up new avenues for protein design using machine learning algorithms.