Published in final edited form as:
J Phys Condens Matter. 2007 July 18; 19(28): 285220–285230.

Shape-dependent designability studies of lattice proteins

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

[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.

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.

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}(˜3.2×10^{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.

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 $\langle R_g^2 \rangle$ for each shape was computed by using the formula:

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

where N is the number of nodes ($N = 24$) , and $\mathbf{r}_{\mathrm{i}}$ 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.

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 inFigure 6. There are 3269 and 4752different sequences that fold to these most designable structures. The conformation A inFig. 6belongs to the shape # 12, 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. 6belongs 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.

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.

Reference List

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. [PubMed: 2385597]
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. [PubMed: 2334692]
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. [PubMed: 10061433]
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(34): 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):50–54.

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. [PubMed: 8662562]

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. [PubMed: 9560215]

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. [PubMed: 10368303]

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. [PubMed: 11381527]

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 CT, Su ZY, Gwan JF, Hao BL, Hsieh CH, Lee HC. Mean-field HP model, designability and alpha-helices in protein structures. Physical Review Letters 2000;84:386–9. [PubMed: 11015917]

34. Shih CT, Su ZY, Gwan JF, Hao BL, Hsieh CH, Lo JL, 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(12): 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. [PubMed: 7877968]

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. [PubMed: 8346235]

41. Shakhnovich EI. Proteins with Selected Sequences Fold Into Unique Native Conformation. Physical Review Letters 1994 Jun 13;72(24):3907–10. [PubMed: 10056327]

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. [PubMed: 1570343]

43. England JL, Shakhnovich EI. Structural determinant of protein designability. Physical Review Letters 2003;90:218101. [PubMed: 12786593]

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. [PubMed: 15741509]

45. England JL, 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. [PubMed: 12843403]

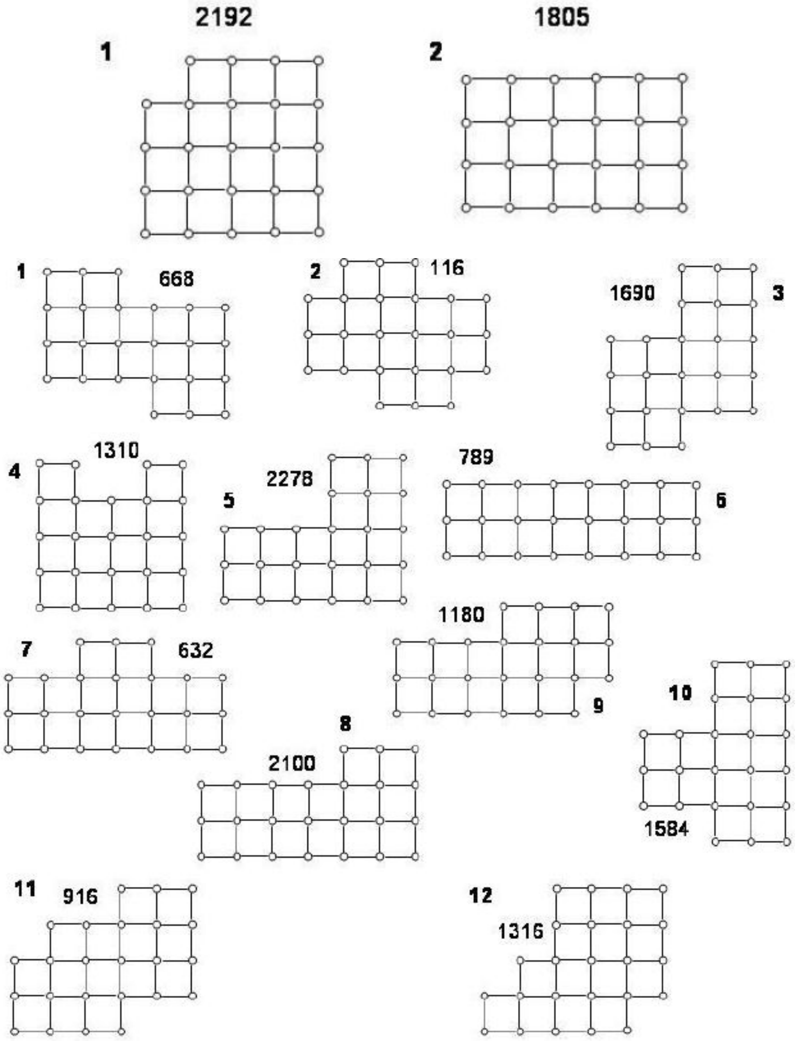


Figure 1.

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.

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.

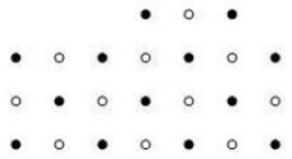
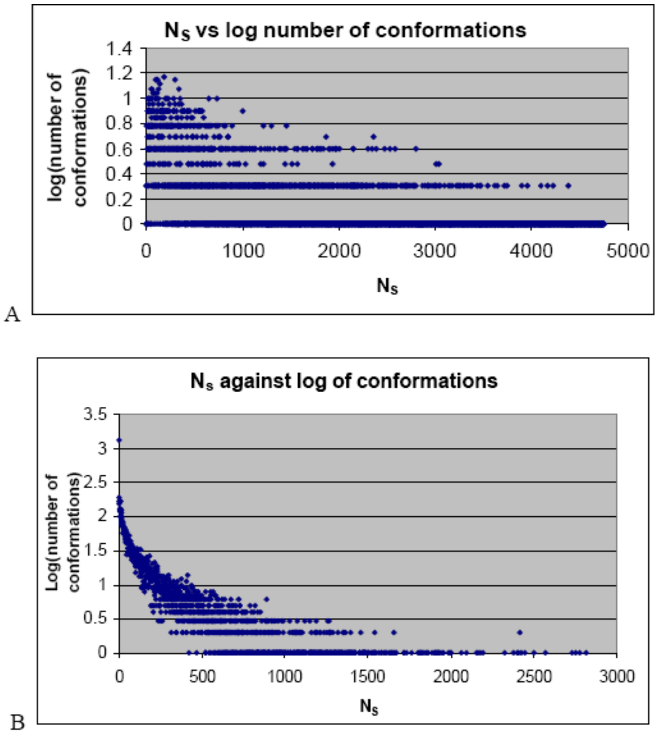


Figure 2.
A shape that is impossible to fill completely with a Hamiltonian path or a circuit. Black and white nodes illustrate chessboard-like feature of the square lattice. Growing a chain will leave unoccupied nodes in all cases.



B

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.

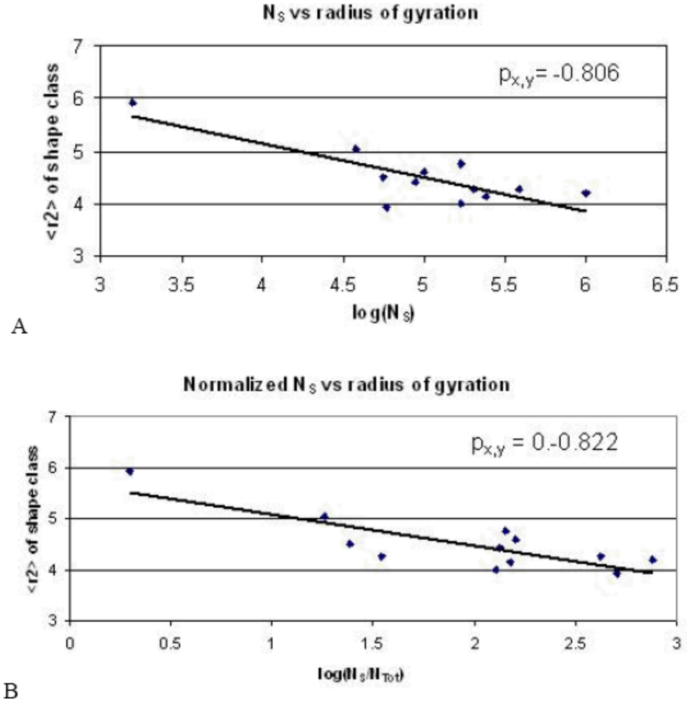


Figure 4.

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.

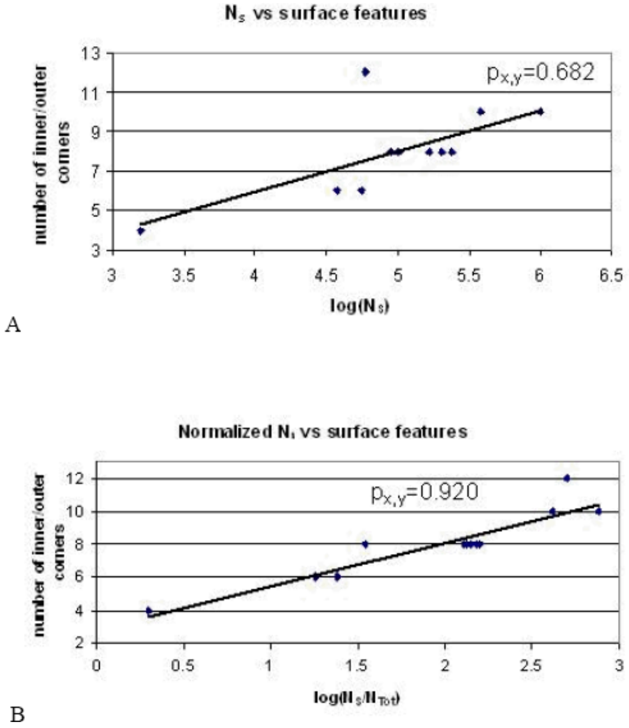


Figure 5.

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.

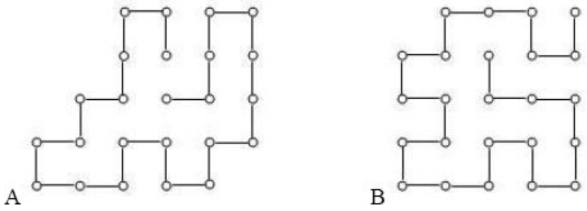


Figure 6.

A-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.

Table 1A

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 2B

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