Published in final edited form as:

Proteins. 2012 June ; 80(6): 1683–1693. doi:10.1002/prot.24067.

Comparative Modeling and Protein-Like Features of HP Models on a Two-Dimensional Lattice

Sergio Moreno-Hernández^{1,2} and Michael Levitt^{2,*}

¹Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305

²Department of Physics, Stanford University, Stanford, California 94305

Abstract

Lattice models of proteins have been extensively used to study protein thermodynamics, folding dynamics and evolution. Our study considers two different hydrophobic-polar models on the two-dimensional square lattice: the purely hydrophobic-polar (HP) model and a model where a compactness-favoring term is added. We exhaustively enumerate all the possible structures in our models and perform the study of their corresponding folds, HP arrangements in space and shapes. The two models considered differ greatly in their numbers of structures, folds, arrangements and shapes. Despite their differences both lattice models have distinctive protein-like features: (1) Shapes are compact in both models, especially when a compactness-favoring energy term is added. (2) The residue composition is independent of the chain length and is very close to 50% hydrophobic in both models, as we observe in real proteins. (3) Comparative modeling works well in both models, particularly in the more compact one. The fact that our models show protein-like features suggests that lattice models incorporate the fundamental physical principles of proteins. Our work supports the use of lattice models to study questions about proteins that require exactness and extensive calculations, such as protein design and evolution, which are often too complex and computationally demanding to be addressed with more detailed models.

Keywords

lattice models; self-avoiding walk; residue composition; hydrophobicity; protein-like; protein universe

INTRODUCTION

Computer simulations provide insights into proteins to compliment the experimental information and increase our theoretical understanding. Protein modeling is challenging due to the tradeoff between the complexity of the models and the computational resources available. The more complex the model, the larger its number of degrees of freedom and the more resources required. At the high extreme of the complexity scale lie very detailed *all-atom* models that account for all the atoms explicitly (including solvent molecules) and use sophisticated interaction energies. They have been used for example to simulate folding explicitly^{1,2}. On the lower extreme of the complexity scale lie *lattice* models^{3,4} which are minimalistic representations of proteins as chains on a regular lattice with very simple interaction energies. These models have also been used to simulate folding, among other applications. There are approaches of intermediate complexity, such as rotamer libraries for

*Correspondence to: Michael Levitt, Department of Structural Biology, 299 Campus Dr West, Stanford, CA 94305. michael.levitt@stanford.edu. Phone: 650-725-0754. Fax: 650-723-8464.

side chains⁵, discrete residue conformations⁶, protein fragment libraries^{7–9} and united-atom models¹⁰. In practice, computer simulations of proteins can involve multiple stages combining models of different complexities. For example, protein structure prediction methods¹¹ often use simplified models in the early stages and finish with all-atom models resembling the experimental crystal structures. Among the different approximations for protein models, our work focuses on the lattice models.

Lattice models are so termed because they represent proteins as a chain of beads (one or more beads per amino acid, although side chains^{12–14} can also be incorporated) whose allowed conformations are self-avoiding walks on a regular lattice. The most commonly used lattices are the simple cubic^{15–19} in three dimensions and the square in two dimensions^{17,20–24}. Other lattices can be considered, like the tetrahedral^{25–27}. Naturally occurring proteins have twenty types of amino acids, however as lattice models are inherently very simple, smaller alphabets with fewer residue types are often considered. For example, Hydrophobic-Polar (HP) lattice models²⁸ consider only two types of residues, hydrophobic (H) and polar (P), with interaction energies accounting for the hydrophobic effect, which is thought to be the driving force for folding^{29,30}. In particular, one of the most popular HP models (that we term *purely hydrophobic-polar*) has non-zero interaction energies only for HH contacts^{20,31}. This interaction energy favors H residues to interact with other H residues, instead of with P residues or with the solvent. Other HP models^{24,27,28,32} consider modified interactions but also incorporate the hydrophobic effect. Whereas HP models only have two residue types, models with up to twenty residue types have also been considered^{33–35}, with interaction energies given by the Miyazawa-Jernigan matrix³⁶, which is derived from the structures of real proteins. Remarkably, all these models are mostly qualitatively similar, likely because they all incorporate the hydrophobic effect³⁰. Furthermore, lattice models have given insights into the energetic interactions in real proteins^{28,37–39}.

The applications of lattice models are rich and varied, for example protein design. Folds differ greatly in their designabilities in nature^{40–43} and in lattice models^{17,27,34,35}. A correlation between designability and stability has been shown in lattice models, with the more designable folds being more thermodynamically stable^{17,30}. Proteins with sequences similar to the profile of the fold tend to be more stable and fold faster than those proteins with the same fold but with sequences further from the profile^{23,24,44}. Possible relationships between symmetry and fold designability have also been investigated^{17,45}. Lattice models also play an important role in the study of the protein folding process. For example, different sequences are designed to have a certain fold as their lowest energy conformation and Monte Carlo simulations are performed to study their folding. It was found that the proteins under study performed a random conformational search until a specific nucleus was formed and after that folding proceeded quickly towards the native state^{18,33}. Others found a similar behavior, however no specific folding nucleus was formed and folding proceeded along multiple pathways^{46–48}. Lattice models have been used to study the folding pathways^{12,22,25,49,50}, the connection between folding rate and native-state energy gap^{16,31,37,46}, the relationship between kinetics and thermodynamics^{51,52}, and cooperativity^{13,21,53–56}. In particular, it was found from kinetic and thermodynamic studies that HP lattice models in two dimensions do not fold as cooperatively as real proteins since their energy level distribution does not accurately represent that of proteins^{55,56}, although cooperativity increases if three dimensions, extra energy terms and side-chains are considered^{13,53}. Finally, lattice models are used in evolutionary studies too^{44,57} since they provide a well defined and unique relationship between sequence and structure. All the lattice proteins with the same fold are assumed to perform the same function and evolution is simulated making random mutations on a starting population of proteins and letting the

system evolve with time^{23,24,35,44,58,59}. Recombination events have also been included to study the relative roles of mutation and recombination^{60,61}.

Lattice models have played very important roles to understand real proteins in terms of their designability, stability, folding and evolution. As mentioned above, these models can be used in conjunction with more detailed ones and for certain applications they are still needed. The goal of our study is to examine the protein-likeness of lattice models in characteristics such as their compactness, residue composition and comparative modeling behavior. We do this by considering two different HP lattice models and also analyzing real proteins for proper comparison.

MATERIALS AND METHODS

Two Models Considered

In our study we consider two hydrophobic-polar models on the square lattice. We use a very simple residue alphabet with only two types of residues: hydrophobic (H) and polar (P). Residue interactions are pairwise, with a given residue interacting only with its closest neighbors not adjacent along the chain (these contacts are excluded as a residue has the same adjacent residues along the chain regardless of the walk). The interaction energy can only take three values: E_{HH} , E_{HP} and E_{PP} depending if the contact is of type H-H, H-P or P-P, respectively. We consider two hydrophobic-polar models that differ in the triplet of energy terms (E_{HH}, E_{HP}, E_{PP}). The first model, which has $(E_{HH}, E_{HP}, E_{PP}) = (-1, 0, 0)$, is named *HP100* and is the purely hydrophobic-polar model mentioned in the introduction. The second model, which has $(E_{HH}, E_{HP}, E_{PP}) = (-2, -1, -1)$, is named *HP211*. It is derived from the first one by adding a favorable energy of -1 , the *compactness-favoring* term²⁴, to each contact formed regardless of the types of the interacting residues.

Data Generation by Exhaustive Enumeration

For a residue alphabet of size A and chain length m there are A^m possible sequences. Considering only two residue types instead of twenty reduces the number of sequences greatly, but it still grows exponentially as 2^m . For some applications sampling of the sequences^{15,17,38,45,62} suffices, however exhaustive enumeration is generally performed when feasible^{17,19,63,64}. Real proteins generally have a single conformation for which the free energy is lowest. This guarantees that under physiological conditions a protein will be in the right conformation and therefore perform its biological function properly. By analogy with real proteins and given the very different degeneracies of the lowest-energy states in lattice models^{20,63} it is common to consider for study only those sequences with a single lowest-energy conformation^{16,19,31,38,60,61} or those for which there is only one conformation within a certain energy gap from the lowest energy level^{24,33,46}. In principle, finding the native state for an arbitrary sequence requires calculating its energy on every possible walk on the lattice. This is challenging as in the two-dimensional square lattice the number of walks grows quickly with the chain length m as $\sim 2.64^m$ ^{28,64}. All the walks can be considered explicitly for shorter chains^{27,38,63,64}. However, it is more common to consider only those walks that fit a particular compact shape^{17,19,35,38,62,65}, reducing dramatically the size of the conformational space and allowing the study of longer chains. This restriction is reasonable since real proteins are compact, however it might introduce artifacts^{16,28,64}.

The top part of Fig. 1 illustrates our data generation process. For a given chain length m and HP model we consider all HP sequences and all walks on the two-dimensional square lattice. We only consider a sequence for analysis if it has a unique lowest-energy state and term such a sequence a *folding sequence*. We term the walk corresponding to its native state a *fold*. Together a folding sequence and its fold define a *structure*. We find all the folding

sequences and all the folds and together they determine the data for our study: the set of all the structures. We perform these calculations for both the HP100 and HP211 models for chain lengths up to 25.

Our approach is very computationally demanding. Taken together, exhaustive consideration of all HP sequences and all walks requires about $2^m \times 2.64^m = 4.64^m$ calculations, exponentially growing with m . We can perform all calculations explicitly for chains up to 19 residues long, however for longer chains a different approach is required. Sophisticated methods have been devised to speed up calculations of this type in special cases^{19,64,66–68}. The speedups achieved are remarkable, however exhaustive calculations for long chains and arbitrary energy terms still remain intractable. To run our calculations for chains of lengths 20 through 25, we considered all the walks of length m with a number of contacts above a conservative threshold and separated them into blocks of similar sizes. For each sequence runs were performed in parallel, obtaining the degeneracy of the lowest energy level considering all the walks in each block. Once all the runs were performed, we combined their outputs to obtain the set of all the folding sequences. Next we repeated all these calculations with a lower value of the threshold for the number of contacts. We kept lowering the threshold until we observed no change in the set of folding sequences and their corresponding folds (see Table S2 in the Supporting Information for the required thresholds). Note that our approach is straightforward to implement and works independent of the choice of energy terms. Although not as elegant or efficient as the optimization methods mentioned before, it is parallelizable so chain lengths up to 25 were computed, the longest ones reached in previous exhaustive unconstrained studies of this kind⁶⁴. In addition the Irback & Troein method⁶⁴ cannot be used for our HP211 model.

Data Analysis

We consider for analysis the sequence, fold, arrangement and shape corresponding to each of the structures generated, as illustrated in the lower part of Fig. 1. The definitions of these objects are very intuitive. Representatives are picked so they are not related to the others by trivial transformations (rotations, translations and mirror-images). However, by analogy with real proteins, we consider that the chain ends are distinguishable. Therefore, for sequences we consider a folding sequence and its reverse as different^{19,63,64}. The folds of both sequences are superimposable, with one sequence starting at one end and the reversed sequence starting at the other end¹⁹. As in the case of sequences, we also consider that a fold and its reverse are different, with distinguishable chain ends. Furthermore, both folds are adopted by the same number of folding sequences, as the sequences that take one of the folds are identical upon reversal to the sequences taking the other fold; this occurs because our 2D lattice has no chirality.

RESULTS

Structures, Folds, Arrangements and Shapes

We start characterizing our models by finding the number of structures, folds, arrangements and shapes, as shown in Fig. 2A. The numbers of naturally occurring structures and folds are still unknown, although the number of folds has been estimated to be about 1,000^{69–71}. On the other hand, since our results are exhaustive, we can answer questions of this type for the universe of simple proteins considered here. The number of structures in our models is between two and four orders of magnitude larger than the number of shapes. For example, for $m=25$ there are 765,147 structures⁶⁴ and 5,080 shapes for HP100; and 4,329,563 structures and 614 shapes for HP211. The number of folds and arrangements lie in between those of structures and shapes. For HP100 they are comparable to each other, for example for $m=25$ there are 107,336 folds⁶⁴ and 160,722 arrangements. For HP211 the number of

folds is always smaller than the number of arrangements (with the slight and minor exception of $m=14$), for example for $m=25$ there are 66,154 folds and 309,643 arrangements. In terms of orders of magnitude, for both models it holds that $\#structures > \#folds \sim \#arrangements > \#shapes$. When comparing the two models there is a reversal of trends: while the HP100 model has *fewer* structures and arrangements than the HP211 model, it has *more* shapes. For folds the numbers are *similar*, with HP100 having slightly fewer folds than HP211 for shorter chains ($m<18$) and more folds for longer chains.

Percentage of Folding Sequences

Fig. 2B shows the percentage of HP sequences that are folding sequences for different chain lengths. This is an important characteristic of our models as it indicates the chance that a sequence is a folding sequence and is therefore considered for analysis. For HP100, about 2.4%^{28,63,64} of the sequences are folding sequences, regardless of the chain length. For HP211, about 12.6% of the sequences are folding sequences, also independent of the chain length. Overall, in HP211 there are about six times more structures than in HP100. However, for HP211 the values fluctuate greatly with the chain length, m . These variations with the chain length favoring certain values of m are termed *magic numbers*⁷² and occur because m is discrete. As shown throughout the paper, this phenomenon is more prominent in HP211.

Maximum Fold Family Size

Fig. 2C shows the maximum fold family size as a function of the chain length. We consider that a fold family encompasses all the folding sequences that take that fold, so the size of a fold family is the number of folding sequences with that particular fold. The fold with the largest family is the most designable one (by definition), has enhanced stability and is particularly important for evolutionary reasons^{17,30,35,44,57,58,73}. The plot shows that the size of the largest fold family grows exponentially with the chain length. The fluctuations, which are due to the occurrence of magic numbers, are larger for HP211 as seen with the percentage of folding sequences. Notably, in HP211 the maximum fold family size is always considerably bigger than in HP100, which has important implications for comparative modeling, as will be shown later. For example, for $m=25$ the maximum family size is 326 in the HP100 model⁶⁴ whereas it is 8,348 in the HP211 model.

Most popular Folds, Shapes & Arrangements

Fig. 3 depicts the most popular folds, shapes and arrangements in both models for $m=25$ labeled by the number of structures that adopt each, providing a visual summary of our models. We define the *popularity* of a fold, arrangement or shape as the number of structures that have that particular attribute. Note that popularity generalizes the concept of family size that we used before in the case of folds. The most popular folds are of particular interest due to their enhanced stability and their important role in evolution^{17,30,35,44,57,58,73}. The most popular folds in the HP100 model look very similar to each other; for example the three most popular folds have three strands that run diagonally. The most popular folds in the HP211 model also look very similar to each other; for example they all have replicas of a substructure resembling the head of a hammer. Therefore the most popular folds in a given model look alike; however the most popular folds in HP100 look very different from the most popular folds in HP211.

The most popular shapes are very compact in both models. The most compact shape that can occur with $m=25$ is a perfect 5×5 square. This is in fact the most popular shape in the HP211 model, whereas it ranks 283rd in terms of popularity in the HP100 model and therefore is not shown. In HP100 the most popular shapes are very compact but not as compact as possible. In HP211 the five most popular folds take the most popular shape. For HP100 the situation

is very different: the fourth most popular fold takes the most popular shape whereas the rest of the HP100 folds depicted take shapes that are not even shown (the folds ranking 1, 2 and 3 take the 35th ranking shape and the fold ranking 5 takes the 9th ranking shape). This illustrates that the most popular folds do not necessarily take the most popular shapes, particularly in the HP100 model.

For the most popular arrangements the situation is similar in terms of their shapes. In HP211 all the arrangements depicted take the 5×5 shape, the most popular one. On the other hand in HP100 each of the arrangements depicted takes a different shape, none of which is the most popular one. Arrangements show the distribution of the HP residues in space, determining the patterns of hydrophobic and polar burial and exposure. All the arrangements shown have completely hydrophobic cores. On the other hand their surfaces are not completely polar i.e. they all have some hydrophobic residues exposed. Arrangements with perfectly hydrophobic cores and perfectly polar surfaces do also occur in our models but they are much less popular and therefore not shown in the figure. For example, the most popular arrangement in HP211 is adopted by 2,578 structures whereas the arrangement with the same shape, an all-polar surface and an all-hydrophobic core is adopted by only 369 structures. Partial hydrophobic exposure makes arrangements more popular.

Compactness

Next we compare our lattice models to real proteins in terms of their compactness, as shown in Fig. 4. The radius of gyration, R_g , gives the typical size of a structure. To remove the chain length dependence and compare with real proteins, we measure the compactness by a normalized radius of gyration, R_g' , defined as $R_g/\text{Min}(R_g)$. The closer R_g' is to 1.0, the more compact a structure is. Compactness of lattice structures has been typically measured in terms of their number of contacts⁷². This is equivalent to our compactness measure, R_g' , as the more contacts a structure has, the smaller its radius of gyration⁶³. Fig. 4A shows the compactness in our lattice models as a function of the chain length m . The solid lines show that in both models structures are on average very compact with R_g' close to 1.0. Comparing the two solid lines we find that the compactness-favoring term indeed makes structures in HP211 more compact than in HP100. As shown by the standard deviation bars, the variation in compactness is very small, particularly in HP211. The dashed lines show that increasingly less compact (but still compact) structures occur for longer chains, particularly in HP100. In summary, both models produce compact structures, particularly HP211.

Fig. 4B shows the compactness of real proteins as a function of the chain length. Most single-domain proteins in nature have a globular shape and are therefore compact, as shown by the solid line being close to 1.0. For our analysis we consider a representative set of proteins from the PDB and our results agree with related studies^{74–76}. We find compactness is on average almost independent of the chain length, just like we find in our lattice models. Note that proteins in the PDB interact with each other in the crystal, introducing artifacts in the dashed line showing the maximum R_g' . A detailed quantitative comparison between our lattice models and real proteins is beyond this study, however qualitatively structures in both the HP100⁶³ and HP211 models are indeed compact like most real single-domain globular proteins.

Residue Composition

We aim to study how protein-like our lattice models are in terms of their residue composition, as shown in Fig. 5. We measure the hydrophobicity of a sequence by the percentage of residues that are hydrophobic, % H . Fig. 5A shows the hydrophobicity in our lattice models as a function of the chain length, m . The solid red line shows that in HP100

the average hydrophobicity is about 54% independent of the chain length^{64,77}. The solid blue line shows that in HP211 the average hydrophobicity is about 50%, also independent of m . The standard deviation bars show that the typical variation in hydrophobicity in our models is about 9.0%. The dashed lines show that the hydrophobicity in our lattice models adopts a very wide range of values, of about 20–80%, being slightly wider for HP211. The range increases with m since the minimum number of hydrophobic and polar residues is almost independent of m (see Table S3 in the Supporting Information). In summary, our two models behave very similarly in terms of their residue composition, with HP100 being slightly more hydrophobic.

Fig. 5B shows the hydrophobicity in real proteins as a function of the chain length. Since real proteins have 20 types of amino acids and our lattice proteins only have two types of residues, we partition the naturally occurring amino acids into the hydrophobic and polar categories in order to make a comparison. The plot shows that in real proteins the average hydrophobicity is about 51%, independent of m . As shown by the bars, the typical variation in hydrophobicity is about 4.5%, half of that for our lattice models. The dashed lines show that the percentage of hydrophobic residues lies in the range 35–65%, which is significantly smaller than that in our lattice models. The smaller spread and range of the hydrophobicity in real proteins might be due to aggregation avoidance as proteins with a high content of hydrophobic or polar residues may aggregate and not be able to perform their biological function. However, in our lattice models aggregation cannot occur as sequences fold in isolation. In summary, being about 50% hydrophobic independent of the chain length for the range of lengths considered ($m=12$ to 25), our lattice structures are protein-like in terms of their residue composition.

Comparative Modeling

Comparative modeling is one of the most important and commonly used tools for protein structure prediction⁷⁹. It consists in predicting the fold of a query sequence using as template the fold of a target protein whose sequence is similar to that of the query. Comparative modeling works because, in general, similar sequences have similar folds. So far we have investigated how protein-like our lattice models are in terms of their compactness and residue composition. To study how well comparative modeling works in our lattice models, we consider the following simple yet realistic scenario. Given a query sequence we predict its fold using as targets those sequences with the same length m and with a fixed number d of mutations away from the query. We measure the Sequence Identity, *SeqID*, between the query i and a target j in percent units by the formula $SeqID(i,j)=100(m-d)/m$. As done elsewhere in our work, we only consider sequences that are folding sequences (see Fig. 1); therefore each query sequence has a corresponding fold, that we term its *true* fold. We assign a fold to the query using three different methods, as explained below. Once a fold is proposed for the query we deem the prediction *correct* only if the predicted fold is *exactly* the same as the true fold of the query. Since in our lattice models we know the true fold of all the query sequences, we can easily determine if a prediction is correct or not. We assess our predictions by the Prediction Accuracy, namely, the fraction of predictions that are correct. Fig. 6 shows the Prediction Accuracy as a function of the Sequence Identity for $m=23$ in the HP100 and HP211 models for our three comparative modeling scenarios, broken down by ranges of the fold family size of the queries.

The first comparative modeling method picks one of the targets randomly and assigns its fold to the query. We term this approach the **Random** method and it provides a baseline for the other two methods. To characterize the resulting Prediction Accuracy vs. Sequence Identity curve with a single parameter, we use the Identity Threshold, defined as the

Sequence Identity above which the Prediction Accuracy is greater than 0.5. The Identity Threshold provides a simple way to assess and compare our three methods and to gain insight into how comparative modeling works in our lattice models: the lower it is, the better comparative modeling works. As shown in Fig. 6A, the Random method produces poor results, with Identity Thresholds of about 90% in both lattice models. The curves look slightly better for queries belonging to popular folds and for the HP211 model.

The second comparative modeling method picks the most popular of the folds of the targets and assigns it to the query. We term this approach the **Majority** method and the resulting plot is shown in Fig. 6B. Note that choosing the most popular fold is like clustering the targets by fold and picking the largest cluster. In both models, the more popular the fold of the query, the better the curves look. For queries belonging to the least popular folds the Identity Threshold in both models is about 90%, and the curves resemble those in the Random method. For queries belonging to the most popular folds, the Identity Thresholds are significantly lower, about 75% in HP100 and 70% in HP211. Therefore the Majority method works better than the Random method, particularly for the HP211 model.

The third comparative modeling method threads the query sequence on the folds of the targets, picks the most energetically favorable one and assigns it to the query. We term this approach the **Threading** method and the resulting plot is shown in Fig. 6C. This method uses the energy terms of the corresponding lattice model to calculate the energy of the query sequence when threaded on the folds of the targets and picks the fold with the lowest energy. In this framework threading⁸⁰ has a simple and insightful interpretation: the prediction will be correct only if the true fold of the query is among the folds of the targets. This is because if the true fold is among the folds of the targets, the query sequence will have the lowest energy when threaded on its true fold, which is therefore picked. If the true fold is not among the folds of the targets, the prediction will not be correct no matter which fold is picked. In both models, the more popular the fold of the query, the better the curves look. In particular, the curves for queries belonging to the least popular folds look significantly better than the corresponding ones in the Random and Majority methods. For queries belonging to the most popular folds the Identity Thresholds are the lowest, about 65% in HP100 and 45% in HP211. Therefore the Threading method works the best, particularly for the HP211 model.

DISCUSSION

In this paper we consider two HP models and study their protein-like features. Our study exhaustively enumerates all the structures in the HP100 model and, as referenced throughout this text, complements previous findings in that system^{63,64}. To make sure that our results are not biased by one choice of the energy terms we consider a second model, HP211. Similarly, instead of considering a single value for the chain length, we consider all values up to 25 and find consistency in the trends, despite the magic-number patterns. A two-dimensional lattice was preferred over a three-dimensional one, as the ratio of the numbers of buried and exposed residues for the chain lengths considered is closer to that in real proteins^{17,20,21}. To avoid sampling artifacts, we exhaustively obtain all the structures and characterize our models by their number of structures, folds, arrangements and shapes. We find that the choice of energy terms, which is the difference between our models, affects the numbers of these objects. In particular, adding a compactness-favoring term reduces the number of shapes and increases the number of structures in our protein universe. This leaves as an open question what the number of naturally occurring proteins and folds would be if the naturally occurring amino acids were different, or more broadly how to design different protein universes.

Our lattice models are protein-like in terms of comparative modeling, which is at the core of the protein structure prediction field. We use three different methods to predict the folds of the query sequences. The Random method is simply a baseline and both the Majority and Threading methods work reasonably well, especially Threading. Comparative modeling works better for queries belonging to popular folds because, given a query sequence, it is easier to find another member of the same fold family to use as a template. For this reason comparative modeling works better in HP211, as in this model folds can be much more popular than in the HP100 model (see Fig. 2C). Furthermore, in our simple framework threading is an excellent method for comparative modeling and we provide a simple explanation for why it works. For real proteins the full power of threading is limited since the energy terms are not yet fully understood. Our results suggest that developing more accurate energy terms will improve comparative modeling, particularly threading techniques. Our lattice models are protein-like also in more straightforward ways, namely in terms of their compactness and residue composition. Structures in both models are very compact, like most real single-domain globular proteins. Importantly, compactness in our models arises without imposing any restrictions on the walks considered. Adding a general attractive term makes structures in the HP211 model more compact, as expected. However, unexpectedly, this also makes folds more popular, which in turn makes comparative modeling work better. Our lattice models also resemble real proteins in terms of their residue composition, with about half of their residues hydrophobic.

Thanks to their simplicity, lattice models require the least computational resources, permitting the most exhaustive enumeration. Furthermore, since all their parameters are known and the calculations can be performed thoroughly, their results are exact and derived rigorously from few principles. Detailed protein models, on the other hand, often require too many computational resources and the underlying physical principles of naturally occurring proteins they try to capture are still not fully understood. For these reasons, lattice models provide excellent (and sometimes the only) systems to address some of the most elusive yet important questions about proteins. These are questions for which there is not enough experimental data or too hard to study extensively with more detailed models, such as protein design, evolution and the nature of the protein universe⁸¹ (note that two complete universes were obtained in this work, one for each model). Therefore, lattice models have been successfully applied in very different areas despite their limitations, most notably their insufficient cooperativity. Lattice models consider only a few proteins features, namely the chain self-avoidance, the hydrophobic effect and the uniqueness of the native states. The fact that our models are protein-like in terms of their compactness, residue composition and comparative modeling indicates that in real proteins these properties probably arise from the simple principles considered by lattice models, suggesting that lattice models succeed at capturing most of the essential features of proteins. In summary, our study supports the validity and use of lattice models of proteins when exact extensive calculations are needed and specific details can be ignored.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health Award GM041455 to M.L. Computations were performed on Stanford's Bio-X² supercluster (National Science Foundation Award CNS-0619926). S.M. was supported by a Stanford Bio-X fellowship. Michael Levitt is the Robert W. and Vivian K. Cahill Professor of Cancer Research. We thank Dr. Golan Yona for feedback on the manuscript.

References

1. Petrone PM, Snow CD, Lucent D, Pande VS. Side-chain recognition and gating in the ribosome exit tunnel. *Proceedings of the National Academy of Sciences of the United States of America*. 2008; 105(43):16549–16554. [PubMed: 18946046]
2. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, Wriggers W. Atomic-level characterization of the structural dynamics of proteins. *Science*. 2010; 330(6002):341–346. [PubMed: 20947758]
3. Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS. Principles of protein folding -- A perspective from simple exact models. *Protein Sci*. 1995; 4(4):561–602. [PubMed: 7613459]
4. Kolinski A, Skolnick J. Reduced models of proteins and their applications. *Polymer*. 2004; 45(2): 511–524.
5. Ponder JW, Richards FM. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology*. 1987; 193(4): 775–791. [PubMed: 2441069]
6. Park BH, Levitt M. The complexity and accuracy of discrete state models of protein structure. *J Mol Biol*. 1995; 249(2):493–507. [PubMed: 7783205]
7. Unger R, Harel D, Wherland S, Sussman JL. A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins*. 1989; 5(4):355–373. [PubMed: 2798411]
8. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*. 1997; 268(1):209–225. [PubMed: 9149153]
9. Kolodny R, Koehl P, Guibas L, Levitt M. Small libraries of protein fragments model native protein structures accurately. *J Mol Biol*. 2002; 323(2):297–307. [PubMed: 12381322]
10. Levitt M. Simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology*. 1976; 104(1):59–107. [PubMed: 957439]
11. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys J*. 2003; 85(2):1145–1164. [PubMed: 12885659]
12. Skolnick J, Kolinski A. Simulations of the folding of a globular protein. *Science*. 1990; 250(4984): 1121–1125. [PubMed: 17840193]
13. Klimov DK, Thirumalai D. Cooperativity in protein folding: from lattice models with sidechains to real proteins. *Fold Des*. 1998; 3(2):127–139. [PubMed: 9565757]
14. Li L, Mirny LA, Shakhnovich EI. Kinetics, thermodynamics and evolution of non-native interactions in a protein folding nucleus. *Nat Struct Biol*. 2000; 7(4):336–342. [PubMed: 10742180]
15. Yue K, Fiebig K, Thomas P, Chan H, Shakhnovich E, Dill K. A test of lattice protein folding algorithms. *Proc Natl Acad Sci USA*. 1995; 92(1):325–329. [PubMed: 7816842]
16. Klimov D, Thirumalai D. Factors governing the foldability of proteins. *Proteins*. 1996; 26(4):411–441. [PubMed: 8990496]
17. Li H, Helling R, Tang C, Wingreen N. Emergence of preferred structures in a simple model of protein folding. *Science*. 1996; 273(5275):666–669. [PubMed: 8662562]
18. Mirny LA, Abkevich VI, Shakhnovich EI. How evolution makes proteins fold quickly. *P Natl Acad Sci Usa*. 1998; 95(9):4976–4981.
19. Cejtin H, Edler J, Gottlieb A, Helling R, Li H, Philbin J, Wingreen N, Tang C. Fast tree search for enumeration of a lattice model of protein folding. *J Chem Phys*. 2002; 116(1):352–359.
20. Lau K, Dill K. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*. 1989; 22(10):3986–3997.
21. Miller R, Danko C, Fasolka M, Balazs A, Chan H, Dill K. Folding kinetics of proteins and copolymers. *J Chem Phys*. 1992; 96(1):768–780.
22. Unger R, Moulton J. Genetic algorithms for protein folding simulations. *Journal of Molecular Biology*. 1993; 231(1):75–81. [PubMed: 8496967]

23. Bornberg-Bauer E, Chan HS. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *P Natl Acad Sci Usa*. 1999; 96(19):10689–10694.
24. Xia Y, Levitt M. Funnel-like organization in sequence space determines the distributions of protein stability and folding rate preferred by evolution. *Proteins*. 2004; 55(1):107–114. [PubMed: 14997545]
25. Skolnick J, Kolinski A. Dynamic Monte Carlo simulations of globular protein folding/unfolding pathways. I. Six-member, Greek key beta-barrel proteins. *J Mol Biol*. 1990; 212(4):787–817. [PubMed: 2329583]
26. Hinds DA, Levitt M. Exploring conformational space with a simple lattice model for protein structure. *J Mol Biol*. 1994; 243(4):668–682. [PubMed: 7966290]
27. Blackburne B, Hirst J. Three-dimensional functional model proteins: Structure function and evolution. *J Chem Phys*. 2003; 119(6):3453–3460.
28. Chan HS, Dill KA. Comparing folding codes for proteins and polymers. *Proteins*. 1996; 24(3): 335–344. [PubMed: 8778780]
29. Dill KA. Dominant forces in protein folding. *Biochemistry*. 1990; 29(31):7133–7155. [PubMed: 2207096]
30. Wingreen N, Li H, Tang C. Designability and thermal stability of protein structures. *Polymer*. 2004; 45(2):699–705.
31. Chan HS, Dill KA. Transition states and folding dynamics of proteins and heteropolymers. *Journal of Chemical Physics*. 1994:9238–9257.
32. Li H, Tang C, Wingreen NS. Are protein folds atypical? *P Natl Acad Sci Usa*. 1998; 95(9):4987–4990.
33. Abkevich VI, Gutin AM, Shakhnovich EI. Specific Nucleus as the Transition State for Protein Folding: Evidence from the Lattice Model. *Biochemistry*. 1994:11.
34. Li H, Tang C, Wingreen N. Designability of protein structures: A lattice-model study using the Miyazawa-Jernigan matrix. *Proteins*. 2002; 49(3):403–412. [PubMed: 12360530]
35. Zeldovich KB, Berezovsky IN, Shakhnovich EI. Physical origins of protein superfamilies. *J Mol Biol*. 2006; 357(4):1335–1343. [PubMed: 16483605]
36. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*. 1985; 18(3):534–552.
37. Unger R, Moulton J. Local interactions dominate folding in a simple protein model. *J Mol Biol*. 1996; 259(5):988–994. [PubMed: 8683600]
38. Noivirt-Brik O, Unger R, Horovitz A. Analysing the origin of long-range interactions in proteins using lattice models. *Bmc Struct Biol*. 2009; 9(1):4. [PubMed: 19178726]
39. Noivirt-Brik O, Horovitz A, Unger R. Trade-off between positive and negative design of protein stability: from lattice models to real proteins. *PLoS Comput Biol*. 2009; 5(12):e1000592. [PubMed: 20011105]
40. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. 1995; 247(4):536–540. [PubMed: 7723011]
41. Orengo CA, Bray JE, Buchan DWA, Harrison A, Lee D, Pearl FMG, Sillitoe I, Todd AE, Thornton JM. The CATH protein family database: a resource for structural and functional annotation of genomes. *Proteomics*. 2002; 2(1):11–21. [PubMed: 11788987]
42. Holm L, Sander C. The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res*. 1996; 24(1):206–209. [PubMed: 8594580]
43. Unger R, Uliel S, Havlin S. Scaling law in sizes of protein sequence families: from super-families to orphan genes. *Proteins*. 2003; 51(4):569–576. [PubMed: 12784216]
44. Chan HS, Bornberg-Bauer E. Perspectives on protein evolution from simple exact models. *Appl Bioinformatics*. 2002; 1(3):121–144. [PubMed: 15130840]
45. Wang T, Miller J, Wingreen N, Tang C, Dill K. Symmetry and designability for lattice protein models. *J Chem Phys*. 2000; 113(18):8329–8336.
46. Sali A, Shakhnovich EI, Karplus M. How does a protein fold? *Nature*. 1994; 369(6477):248–251. [PubMed: 7710478]

47. Klimov DK, Thirumalai D. Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism. *J Mol Biol.* 1998; 282(2):471–492. [PubMed: 9735420]
48. Klimov DK, Thirumalai D. Multiple protein folding nuclei and the transition state ensemble in two-state proteins. *Proteins.* 2001; 43(4):465–475. [PubMed: 11340662]
49. Onuchic J, Socci, Luthey-Schulten, Wolynes. Protein folding funnels: the nature of the transition state ensemble. *Folding and Design.* 1996; 1(6):441–450. [PubMed: 9080190]
50. Pande VS, Rokhsar DS. Folding pathway of a lattice model for proteins. *Proc Natl Acad Sci U S A.* 1999; 96(4):1273–1278. [PubMed: 9990014]
51. Camacho CJ, Thirumalai D. Kinetics and thermodynamics of folding in model proteins. *Proc Natl Acad Sci U S A.* 1993; 90(13):6369–6372. [PubMed: 8327519]
52. Melin R, Li H, Wingreen N, Tang C. Designability, thermodynamic stability, and dynamics in protein folding: A lattice model study. *J Chem Phys.* 1999; 110(2):1252–1262.
53. Kolinski A, Galazka W, Skolnick J. On the origin of the cooperativity of protein folding: implications from model simulations. *Proteins.* 1996; 26(3):271–287. [PubMed: 8953649]
54. Jewett AI, Pande VS, Plaxco KW. Cooperativity, smooth energy landscapes and the origins of topology-dependent protein folding rates. *J Mol Biol.* 2003; 326(1):247–253. [PubMed: 12547206]
55. Chan HS. Modeling protein density of states: additive hydrophobic effects are insufficient for calorimetric two-state cooperativity. *Proteins.* 2000; 40(4):543–571. [PubMed: 10899781]
56. Chan HS, Zhang Z, Wallin S, Liu Z. Cooperativity, local-nonlocal coupling, and nonnative interactions: principles of protein folding from coarse-grained models. *Annu Rev Phys Chem.* 2011; 62:301–326. [PubMed: 21453060]
57. Xia Y, Levitt M. Simulating protein evolution in sequence and structure space. *Current Opinion in Structural Biology.* 2004; 14(2):202–207. [PubMed: 15093835]
58. Tiana G, Shakhnovich BE, Dokholyan NV, Shakhnovich EI. Imprint of evolution on protein structures. *Proc Natl Acad Sci USA.* 2004; 101(9):2846–2851. [PubMed: 14970345]
59. Blackburne BP, Hirst JD. Population dynamics simulations of functional model proteins. *J Chem Phys.* 2005; 123(15):154907. [PubMed: 16252972]
60. Xia Y, Levitt M. Roles of mutation and recombination in the evolution of protein thermodynamics. *Proc Natl Acad Sci USA.* 2002; 99(16):10382–10387. [PubMed: 12149452]
61. Cui Y, Wong WH, Bornberg-Bauer E, Chan HS. Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes. *Proc Natl Acad Sci USA.* 2002; 99(2):809–814. [PubMed: 11805332]
62. Tang C. Simple models of the protein folding problem. *Physica A.* 2000; 288(1–4):31–48.
63. Chan H, Dill K. “Sequence space soup” of proteins and copolymers. *J Chem Phys.* 1991
64. Irback A, Troein C. Enumerating designing sequences in the HP model. *J Biol Phys.* 2002; 28(1):1–15.
65. Covell DG, Jernigan RL. Conformations of folded proteins in restricted spaces. *Biochemistry.* 1990; 29(13):3287–3294. [PubMed: 2334692]
66. Dill KA, Fiebig KM, Chan HS. Cooperativity in protein-folding kinetics. *Proc Natl Acad Sci USA.* 1993; 90(5):1942–1946. [PubMed: 7680482]
67. Yue K, Dill KA. Forces of tertiary structural organization in globular proteins. *Proc Natl Acad Sci U S A.* 1995; 92(1):146–150. [PubMed: 7816806]
68. Backofen R, Will S. A constraint-based approach to fast and exact structure prediction in three-dimensional protein models. *Constraints.* 2006; 11(1):5–30.
69. Chothia C. Proteins. One thousand families for the molecular biologist. *Nature.* 1992; 357(6379):543–544. [PubMed: 1608464]
70. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. *Nature.* 1994; 372(6507):631–634. [PubMed: 7990952]
71. Brenner SE, Chothia C, Hubbard TJ. Population statistics of protein structures: lessons from structural classifications. *Curr Opin Struc Biol.* 1997; 7(3):369–376.
72. Chan HS, Dill KA. Compact polymers. *Macromolecules.* 1989; 22(12):4559–4573.

73. Taverna D, Goldstein R. The distribution of structures in evolving protein populations. *Biopolymers*. 2000; 53(1):1–8. [PubMed: 10644946]
74. Arteca G. Scaling behavior of some molecular shape descriptors of polymer chains and protein backbones. *Phys Rev E*. 1994; 49(3):2417–2428.
75. Ivankov DN, Bogatyreva NS, Lobanov MY, Galzitskaya OV. Coupling between properties of the protein shape and the rate of protein folding. *PLoS ONE*. 2009; 4(8):e6476. [PubMed: 19649298]
76. Hong L, Lei J. Scaling Law for the Radius of Gyration of Proteins and Its Dependence on Hydrophobicity. *J Polym Sci Pol Phys*. 2009; 47(2):207–214.
77. Irback A, Sandelin E. On hydrophobicity correlations in protein chains. *Biophys J*. 2000; 79(5): 2252–2258. [PubMed: 11053106]
78. Creighton ET. *Proteins: structures and molecular properties*. 1993:507.
79. Cozzetto D, Kryshchuk A, Fidelis K, Mout R, Rost B, Tramontano A. Evaluation of template-based models in CASP8 with standard measures. *Proteins*. 2009; 77:18–28. [PubMed: 19731382]
80. Xu J, Li M. Assessment of RAPTOR's linear programming approach in CAFASP3. *Proteins*. 2003; 53 (Suppl 6):579–584. [PubMed: 14579349]
81. Levitt M. Nature of the protein universe. *Proc Natl Acad Sci USA*. 2009; 106(27):11079–11084. [PubMed: 19541617]

Figure 1.

Illustration of the generation and analysis of our data. We consider the two-dimensional square lattice and only two types of residues: hydrophobic (H, in red) and polar (P, in blue). For comparison, we consider two models with different closest-neighbor interaction energies (E_{HH} , E_{HP} , E_{PP}): the so-called *HP100* model with $(-1, 0, 0)$ and the *HP211* model with $(-2, -1, -1)$. **Data Generation (Top):** Given a model and a chain length, m , we enumerated all the HP sequences and all the walks on the lattice. Next we calculated the energy of each sequence on every walk. By analogy with real proteins, a *folding sequence* is a sequence whose energy is lowest on a unique walk, that we term its *fold*. A folding sequence and its corresponding fold together determine a *structure*. We exhaustively found all the structures for both models for m up to 25 i.e. all the sequences that are folding sequences and their corresponding folds. **Data Analysis (Bottom):** For each structure we consider for analysis its Sequence, Fold, Arrangement and Shape. The figure illustrates these objects for a sample structure. The *sequence* of the structure is a folding sequence by definition. The *fold* of the structure is the walk to which its sequence folds i.e. the native state. The *arrangement* of the structure is determined by the residue types and their positions in space regardless of the chain connectivity. The *shape* of the structure is determined solely by the positions in space that are occupied by residues regardless of their types.

Figure 2.

Magnitudes of important quantities characterizing the HP100 and HP211 lattice models.

Panel (A) shows the number of structures, folds, arrangements and shapes as a function of the chain length, m , in the HP100 (red) and HP211 (blue) models. All Counts increase exponentially with m and hence linearly on these log-linear plots (see Table S1 in the Supporting Information for the data with $m=3-25$). **Panel (B)** shows the percentage of HP sequences that are folding sequences as a function of m for both lattice models. On average about 2.4% of all sequences in the HP100 model and 12.6% of all sequences in the HP211 model are folding sequences and therefore associated with a structure. The peaks for HP211 occur at magic number values of m that are the product of two similar size factors, namely

12=3×4, 16=4×4, 20=4×5, 24=4×6 and 25=5×5; this minimizes the perimeter-to-area ratio.

Panel (C) shows the Maximum Fold Family Size (i.e. the number of folding sequences that take the most popular fold) as a function of m for both lattice models.

Figure 3.

Depictions of the five most popular folds (top row), shapes (medium row) and arrangements (bottom row) for chain length $m=25$ in the HP100 (left panel) and HP211 (right panel) models, which give a very vivid picture of our models. Each object is labeled and sorted by its popularity i.e. its frequency of occurrence. Since we consider that the chain ends are distinguishable, for each fold depicted its reverse is also a valid fold that looks identical and has the same popularity (see Methods). Given this repetition, only one of the two folds is shown as a representative.

Figure 4.

Compactness as a function of the chain length, m , for our lattice models and for real proteins. We measure the compactness by the normalized radius of gyration $R_g' = R_g / \text{Min}(R_g)$. **Panel (A)** shows the compactness for the lattice structures in models HP100 (red) and HP211 (blue). $\text{Min}(R_g)$ is computed as the smallest radius of gyration among the structures of a given length. **Panel (B)** shows the compactness for real proteins. We have considered a non-redundant set of 2,401 single-domain proteins from the PDB with lengths between 50 and 200 residues long. As the number of data points is small, $\text{Min}(R_g)$ for a given length is computed from a fitting of the data. In both plots we show for R_g' the mean values (the solid lines joining the circles), the standard deviations (the bars) and the range of values (the dashed lines). In the case of real proteins the data was binned uniformly into 15 bins.

Figure 5.

Hydrophobicity as a function of the chain length, m , for our lattice models and for real proteins. We measure the hydrophobicity by the percentage of residues that are hydrophobic, $\%H$. **Panel (A)** shows the hydrophobicity for the lattice structures in models HP100 (red) and HP211 (blue). **Panel (B)** shows the hydrophobicity for real proteins. We have considered the same set of proteins from the PDB as used to study the compactness (see legend of Fig. 4). In order to calculate the hydrophobicity for real proteins and make a comparison with our lattice models, we have considered the amino acids K, R, D, E, N, Q, S, G and H as polar (P) and T, A, P, Y, V, M, C, L, I, F and W as hydrophobic (H). This partition was done choosing a sensible hydrophobic-polar threshold in the Creighton hydrophobicity scale⁷⁸. In both plots we show for $\%H$ the mean values (the solid lines joining the circles), the standard deviations (the bars) and the range values (the dashed lines). In the case of real proteins the data was binned uniformly into 15 bins.

Figure 6.

Comparative modeling using the Random, Majority and Threading methods in HP100 (red) and HP211 (blue) for chain length $m=23$. The Sequence Identity measures the distance between the query and the target sequences, as the percent of residues that need to be mutated to obtain one from the other. A prediction is considered correct only if the predicted fold is the true fold. The Prediction Accuracy is defined as the fraction of predictions that are correct. We use three different methods to assign a fold to the query sequence given a set of targets. **Panel (A):** In the Random method, one of the target proteins is picked randomly and its fold assigned to the query. **Panel (B):** In the Majority method, the most popular fold among those of the targets is picked and assigned to the query. **Panel (C):** In the Threading method, the query sequence is threaded on the folds of the targets and the most energetically favorable one is picked and assigned to the query. In the last two methods, in the case of ties we choose randomly among the competing candidate folds. In each panel the plots are broken down into three uniformly-distributed ranges of the Fold Family Size of the queries. These plots are obtained with exhaustive sampling. The plots show that comparative modeling works better for the HP211 model, for the Threading method and for queries belonging to big fold families.