# Three-dimensional functional model proteins: Structure function and evolution

Benjamin P. Blackburne and Jonathan D. Hirst[a)]
*School of Chemistry, University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom*

(Received 18 December 2002; accepted 16 May 2003)

The mapping of phenotype onto genotype for a set of functional model proteins is accomplished by exhaustive enumeration on a three-dimensional diamond lattice. Chains of up to 25 monomers are investigated and their evolution characterized. The model is used to investigate the origins of designability. Highly designable functional model protein structures possess contact maps that have a relatively little commonality with other physically allowed contact maps. Although the diamond lattice has the same coordination number as the square lattice, differences between three-dimensional and two-dimensional functional model proteins are observed. One difference is the lower frequency of structures of low designability on the three-dimensional lattice. In other respects, the conclusions drawn from previous studies using the square lattice remain valid in three dimensions. For example, we observe the tendency for longer chains to form larger networks of sequences with greater stability to mutation. We identify various topographical characteristics of the landscapes: evolutionary bottlenecks bridge otherwise unconnected networks, and hub sequences allow rapid movement between the different neutral networks. The diversity of landscapes that arises from even a minimalist model suggests that real proteins have a rich variety of evolutionary landscapes.  [DOI: 10.1063/1.1590310]

## I. INTRODUCTION

The mechanism of protein evolution is natural selection based on function. Function is exhibited as a result of the specific three-dimensional structure or fold of a protein. The fold often includes a binding pocket or active site as a location of activity, and this site is an obvious focus for examining fitness. A mutation may alter the fold of a protein, and thereby the binding pocket, or it may change a residue in the binding pocket, modifying function. Alternatively, a mutation may lie away from the binding pocket and not change the fold, allowing function to be maintained. The fixation of both sorts of mutations leads to two processes. The former process is adaptive evolution, and is the mechanism by which proteins can acquire new and improved function. The latter process is neutral drift; the neutral theory[1] proposes that the most highly adapted proteins are unable to evolve any other way. Understanding the roles of both these processes is an area of current inquiry.

A classic model for visualizing the process of molecular evolution is Wright's fitness landscape.[2] A set of genotypes is connected by point mutations on a landscape, with molecular evolution taking place through successive fixation of mutants, leading to areas of higher fitness. By modeling such landscapes we may hope to answer some evolutionary questions, such as the extent of fixation of neutral mutations in a protein's evolution, and the related question of whether it is possible for a highly adapted protein to improve its fitness.

In order to examine the evolution of proteins in a tractable fashion a variety of models have been proposed. Kauffman[3] employed a method of random assignments of fitnesses in order to model fitness landscapes. The effect of varying the level of epistatic interactions, that is, the dependency of the fitness of one gene on the fitnesses of others, was examined. The organism was represented as a system of $N$ loci. Each locus, or position on the genome, was assigned an allele with a fitness affected by a number ($K$) of other genes. This model, the so-called $NK$ model, has produced a wealth of insights into how the structure of such landscapes may affect evolution, but it does not address the question of how the physical constraints of protein structure affect the landscape structure.

More recently, computer simulations have been used to examine directly how the requirements of fitness and stability affect protein evolution. In one study,[4] molecular dynamics simulations were undertaken for short sequences and the positions of four specific residues appraised for their similarity to a certain active site configuration. Mutation and selection with fitness defined as similarity to the target active site was able to produce proteinlike structures.

However, for a broader view of how molecular evolution works we need to consider vast evolutionary landscapes, and obtaining sufficient experimental data is difficult. The interpretation of genomic data in a form relevant to evolutionary fitness is further complicated by the lack of rapid, reliable protein folding algorithms capable of distinguishing subtle differences between structures. For these reasons, studies frequently employ lattice based models, with a reduced alphabet of amino acids.

Another issue is the definition of fitness. A biological definition of fitness involves success of an organism at reproduction. Applying this definition to a landscape of a gene is difficult, and different approaches are taken, often involving

[a)]Electronic mail: jonathan.hirst@nottingham.ac.uk

 

the stability of the protein. We define fitness as the number of hydrophobes in the binding pocket of the protein. This is an easily characterized physical property that is relevant to a possible function, the nonspecific binding of a hydrophobic substrate.

Coarse-grained models are advantageous in computational studies of proteins, as they restrict conformational space to a level amenable to exhaustive enumeration, while maintaining physical relevance. The use of lattice models in protein folding is well established,[5,6] and their application to evolutionary problems is increasing. In one study, the origins of designability have been examined using a 20-letter lattice model.[7] In another study,[8] a 16-monomer chain on a square lattice was used to model binding to a ligand. The simulation of evolution with selection for compactness compared with selection for binding showed remarkable similarities in the distribution of structures. A two-dimensional lattice model was used to investigate the ability of evolution to produce protein structures that are highly robust to mutation.[9] The natural propensity for evolution to select sequences for robustness to mutation due to the effects of quasispecies[10] was demonstrated. Each sequence is able to contribute mutations to its neighboring sequences, and so the population at each point in sequence space is affected by the population of the surrounding points. Sharp peaks of high fitness cannot be sustained by high levels of mutation from surrounding sequences. This means that wider areas of more moderate fitness are more populated.

After reducing the complexity of conformational space with a lattice model, the size of the landscape is reduced to a level amenable to exhaustive enumeration through the use of a two-letter alphabet such as the HP (hydrophobic/polar) model.[11] This is as simple as possible for a heteropolymer and stems from the physical observation that the hydrophobic interaction is the major determinant in protein folding.[12,13] Each residue is reduced to a single point on the lattice, and is assigned to be either hydrophobic or polar.[11] The power of this concept is illustrated by *de novo* protein design experiments, which select sequences with certain patterns of hydrophobic and polar residues.[14] The success of this ''bury the grease'' strategy indicates the relevance of the HP model to real proteins. The patterning of hydrophobic and polar amino acids can be used to design a sequence to fold to a particular target structure.[15] Furthermore, differential packing of hydrophobic and polar residues is believed to be the origin of the observation that switching two adjacent residues in the protein Arc results in different tertiary and secondary structures.[16] Elements of secondary structures also show distinctive HP patterns. In one study,[17] the HP distributions were shown to be different for parallel and antiparallel beta sheets, with further differences between interior and edge strands.

The ability to search conformational space and sequence space exhaustively makes HP lattice model proteins useful in evolutionary studies, giving insights into the nature of evolutionary landscapes. The existence of a superfunnel topology for neutral nets has been proposed from studies[18] with the HP model. Sequences in a neutral net are shown to be organized such that they are centered around a prototype sequence, with stabilities increasing toward the center. Questions of designability can also be addressed. A minimalist model has been used to demonstrate how highly designable structures are typically the most thermodynamically stable.[19]

Our model is simple, yet rich enough to support a variety of fitnesses. The emergent complexity of the model allows hypotheses about evolution to be tested with different amino-acid alphabets, lattices, and chain lengths, in order to determine their robustness with respect to these conditions. We utilize a three-dimensional (3D) lattice as a step to more realistic structures. Unlike many other studies[20–22] which employ a maximally compact 27-mer on a cubic lattice, restricted to a $3\times3\times3$ cube, we use the diamond lattice. This is a four-coordinate lattice, which reduces conformational space enough to allow exhaustive enumeration of structures up to a moderate length.

We also consider questions of protein designability. Recent work[22] has suggested that two letter models of proteins give rise to a set of designable structures that is different than that produced by larger alphabets, and that conclusions about designability drawn from studies of HP models have limited generality. However, as noted elsewhere[23] this suggestion itself rests on studies of maximally compact $5\times5$ conformations of a 25-mer. Studies which allow all conformations produce fewer sequences that encode for maximally compact structures, and so native maximally compact structures are less designable.

The validity of the HP model on a 3D lattice has also been questioned.[24] It was suggested that the HP model may not display proteinlike characteristics, such as a unique native state with an energy gap and cooperative folding, on a 3D lattice. However, this conclusion is drawn from the study of a set of designed sequences in the HP model on the cubic lattice. Our studies, described in the following, on the diamond lattice using the shifted-HP model[11] show a large number of nondegenerate ground states with energy gaps, which also possess a binding pocket.

## II. METHODS

The HP model assumes a favorable interaction energy when two hydrophobes are in contact on the lattice but are not neighbors in the chain ($E_{\mathrm{HH}}=-1$), while other contacts are neutral ($E_{\mathrm{HP}}=E_{\mathrm{PP}}=0$). This confers a hydrophobic core and polar surface on native structures, as seen in real proteins. However, the inclusion of only attractive and neutral forces tends to lead to highly degenerate maximally compact native states. In contrast to this, real proteins fold into a single native state, often containing a binding pocket which is required for function. In order to accommodate this, repulsive terms can be introduced into the HP model, which leads to the shifted-HP model:[11]

$$E=\begin{pmatrix} -2 & 1 \\ 1 & 1 \end{pmatrix}.$$

In this case, $E_{\mathrm{HH}}=-2$, and $E_{\mathrm{PH}}=E_{\mathrm{HP}}=E_{\mathrm{PP}}=1$.

Previously,[25,26] we have studied the square lattice. In this paper we extend our analysis to the 3D diamond lattice. The diamond lattice is expected to incorporate more realism into

TABLE I. Conformational space of chains on the diamond lattice.

| Number of monomers in chain | Number of conformations | Number of contact maps | Number of nondegenerate contact maps |
|---|---|---|---|
| 19 | 15 094 486 | 39 282 | 8 103 |
| 20 | 43 844 655 | 98 331 | 18 818 |
| 21 | 127 162 522 | 223 676 | 44 461 |
| 22 | 369 043 759 | 569 472 | 112 445 |
| 23 | 1 069 666 168 | 1 302 090 | 234 799 |
| 24 | 3 102 070 729 | 3 307 363 | 642 415 |
| 25 | 8 986 576 978 | 7 647 140 | 1 378 104 |

our model, while maintaining the four coordinate nature of the square lattice, thereby obviating a massive increase in complexity. For the purposes of our enumeration we remove all structures that are duplicates through symmetry or rotation. In Table I, we show how conformational space grows with chain length. The data agree with early enumerations by Wall and Hioe[27] for chains up to length 21. The earlier enumeration ignores symmetry and rotation, so the number of conformations reported earlier is $24\Omega$-12, where $\Omega$ is the number reported as ''number of conformations'' in Table I. The growth in number of conformations is similar to the growth seen on the square lattice;[26] however, the number of contact maps is smaller. This is due to the necessity of making seven moves on the diamond lattice to return to the starting position, rather than the five needed on the square lattice. This means fewer contacts are possible for a given chain length. An example of a chain on the diamond lattice is shown in Fig. 1.

We define a viable sequence as one which satisfies several conditions. It must have a nondegenerate ground state—that is it must possess a single lowest energy conformation. It must have an energy gap—the difference between the energy of the lowest energy conformation and the first excited conformations must be at least two. This will lead to two distinct populations during protein folding, and folding will therefore exhibit a form of two state-cooperativity. Analysis of the 21-mer sequences that were excluded by the energy gap requirement shows that these excluded sequences would exhibit folding that is clearly not cooperative. For example 91% of the structures close in energy to the native structure are ''non-native,'' sharing 50% or fewer of the contacts of the native state. We conclude that it is appropriate to employ the energy gap requirement to exclude sequences that lack an energy gap from further analysis. This criterion does not guarantee that all the functional model proteins fold cooperatively, because this energy gap does not consider whether the structures of the first excited states are similar or dissimilar to the ground state. Structures similar to the ground state will be present in the folding funnel and do not represent a competing energy minimum.[28] Ideally a criterion that is designed to enforce cooperativity should reflect this. However, it may be that there is no energy gap that can capture the level of calorimetric cooperativity shown by real proteins, without the inclusion of nonadditive terms in the energy formulation.[29] Cooperativity may be an evolved feature[30] dependent on side-chain packing and other features.[31,32] With respect to the overall aim of studying models with protein-like phenomenology, the energy gap requirement is useful for excluding model sequences with unproteinlike behavior. More detailed work on cooperativity in lattice models would be desirable, but is beyond the scope of this study.

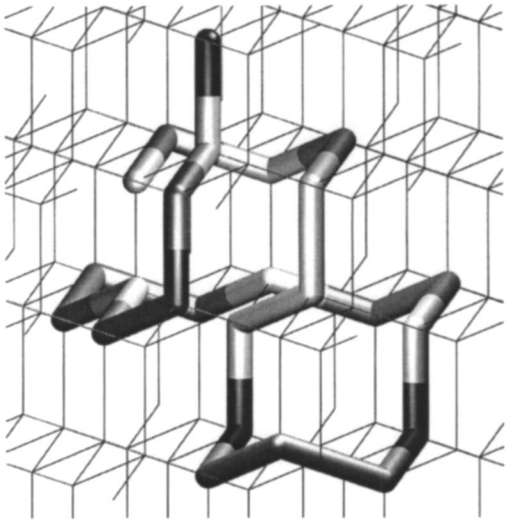

FIG. 1. An example of a 25-mer diamond lattice protein. White residues are hydrophobic, black are polar. Gray residues are subject to variation within the family.

Finally, the lowest energy structure must possess an empty lattice site surrounded by three or four residues. This requirement for a binding pocket is necessary for a protein to be deemed functional. The function of the model protein may then be investigated as a feature distinct from conformation or stability, by characterization of the binding pocket.

For each chain length, we generate all self avoiding random walks on the diamond lattice using the backtracking algorithm described earlier.[25] To evaluate a HP sequence we consider the distribution of its energies over all possible conformations. As previously noted,[11,25,26] we do not need to consider the energy for each conformation explicitly, but merely for each contact map. We can then determine whether a native structure exists for each sequence. To build a complete evolutionary landscape, we consider every possible HP sequence. However, for chains of length 24 and above this becomes infeasible. Instead, we search exhaustively in structure space, but sample sequence space. To obtain our set of candidate sequences for length $n$ monomers we take the set of viable sequences of length $n$-2, and make all possible insertion mutations. To this set, we add all viable sequences of length $n$-1. Finally, we make all possible insertions to this set, to obtain a set of sequences of length $n$. This approach is motivated by the observation that large functional model protein families tend to span a range of lengths through indel mutations.[26] From the viable sequences so found, we evaluate all possible point mutations, until the families have been fully explored. In this way we obtained a set of viable sequences for the 24-mer without exhaustive enumeration of sequence space. Through further insertions to the 23-mers and 24-mers we can find viable 25-mers, from which we can make point deletions to generate further 24-mer candidate sequences. We repeat the process until no more new se-

TABLE II. Characteristics of evolutionary landscapes. The ratio of neutral:adaptive mutations (N:A) and the average number of nonlethal mutations per sequence (M/S) are given.

| Number of monomers in chain | All sequences | | Third largest family | | | Second largest family | | | Largest family | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | N:A | M/S | Size | N:A | M/S | Size | N:A | M/S | Size | N:A | M/S |
| 22 | 1.85 | 1.46 | 31 | 1.48 | 2.00 | 31 | 1.44 | 1.97 | 33 | 1.68 | 1.79 |
| 23 | 2.00 | 1.52 | 31 | 1.48 | 2.00 | 45 | 2.18 | 1.91 | 45 | 2.03 | 2.16 |
| 24 | 2.33 | 1.02 | 34 | 3.56 | 2.15 | 45 | 2.03 | 2.16 | 126 | 2.92 | 2.61 |
| 25 | 2.24 | 1.81 | 86 | 2.52 | 1.97 | 120 | 2.85 | 2.63 | 151 | 2.90 | 2.76 |

quences are found. This process is justified by inspection of the sequences found by exhaustive enumeration: starting with all viable 21-mer sequences, and proceeding with our indel strategy to generate 22-mers and 23-mers, we would find all 23-mer families with at least 25 members, and 55% of all viable 23-mer sequences. The 23-mers not found are in families with fewer than 21 members.

Our functional model proteins are simplified, comprising only hydrophobic and polar residues, and correspondingly we use nonspecific hydrophobic binding as a model of fitness. Previously we defined the fitness, *f*, as the number of hydrophobic residues around the binding pocket.[25,26] On the square lattice, including next nearest neighbors gives eight sites and provides a reasonable range of fitnesses. In order to draw meaningful comparisons between the two lattices we again equate fitness with the number of nearest neighbors and next nearest neighbors (i.e., a total of sixteen possible sites) that are hydrophobic. In practice, the number of occupied sites is lower, rendering a comparable range of fitness for the two lattices.

## III. RESULTS AND DISCUSSION

### A. Neutral and adaptive mutations

A useful measure of the ruggedness of the landscape is the ratio of neutral:adaptive mutations. A flat landscape dominated by neutral mutations would mean that the model proteins are more stable to mutation, while a highly rugged landscape would imply that the fittest sequences are less stable to mutation. Proteins in a highly rugged landscape would have limited evolutionary potential.

Table II shows that the proportion of neutral mutations increases with increasing length and landscape size. Larger landscapes will stabilize a protein with respect to mutation simply because of their size, and also as a result of the increase in the number of neutral mutations. It is less clear how the number of nonlethal mutations per sequence changes with length and how this affects evolution, although there is some propensity for larger landscapes to be more highly interconnected. This implies that larger landscapes exist partly as a result of ''filling in the gaps'' of sequence space, rather than simply an expansion outwards.

A useful measure of ruggedness is the autocorrelation function,[3] which calculates the correlation of fitnesses at different evolutionary distances. Flat landscapes will be highly correlated over long distances. As landscapes become more rugged, fitnesses over large distances will become uncorrelated. A random walk of 2048 nonlethal mutations is taken on the landscape. We then calculate the correlation between the fitness, *f*, at each position, *t*, and that *s* steps away, by considering the covariance divided by the variance:

$$R(t,s)=\frac{E(f_t\times f_{t+s})-E(f_t)\times E(f_{t+s})}{\text{variance}(f)}.$$

This process is repeated 100 times for each landscape and the mean autocorrelations for a selection of landscapes are plotted in Fig. 2.

As we increase the number of steps the autocorrelation decreases in a manner that depends on the structure of the landscape. The majority of the landscapes consist of a single large network containing a homogeneous mixture of fitnesses. These landscapes are rugged and so the autocorrelation falls rapidly, with many landscapes becoming uncorrelated by the third or fifth mutation. While some landscapes exhibit smoothly declining curves, the majority exhibit an oscillatory behavior, with even-numbered steps on the walk increasing autocorrelation and odd-numbered steps decreasing autocorrelation. The difference between the two may be a result of the specific characteristics of the landscapes. Some landscapes are highly rugged, and a step on the walk may be very likely to correspond to a functional mutation, as a hydrophobe in the binding pocket is replaced by a polar resi-

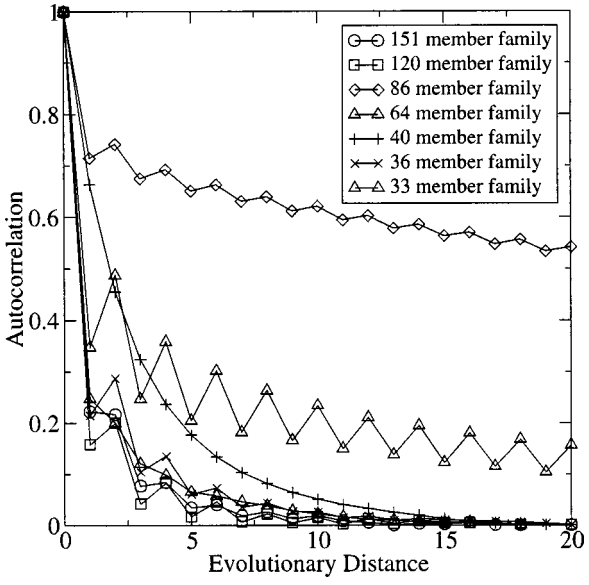


FIG. 2. Autocorrelation of evolutionary landscapes of various families.

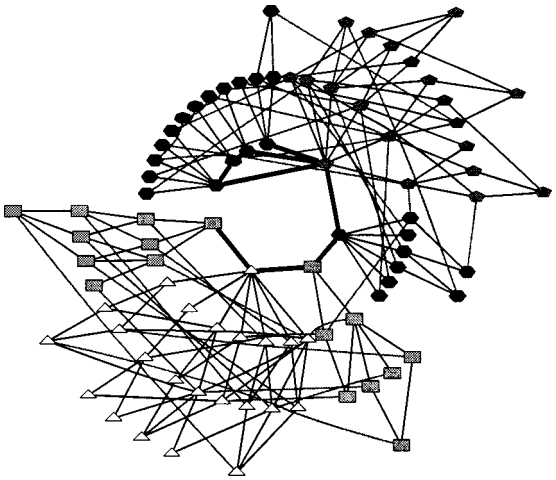

FIG. 3. 86-member family of 25-mers. Nodes correspond to functional model proteins; edges correspond to single point mutations connecting two viable sequences. Hub nodes are closest to the center, connected by bold edges. The number of sides of a node represents the fitness of the sequence (the number of hydrophobes around the binding pocket). Triangular nodes represent sequences with fitness three, nodes with seven sides represent sequences with fitness seven. Darker colors indicate nodes of greater fitness.

due, or vice versa (e.g., the landscape of size 33). The requirement of folding means that landscapes tend to be constrained to a certain number of fitnesses and a subsequent mutation is likely to return the protein to its original fitness, rather than take it further away. The contrasting type of landscape is that exhibited by some of the smaller landscapes (e.g., that of size 40). Rather than being highly homogeneous, they are partitioned, with two fitnesses meeting at an interface. This sort of landscape does not exhibit oscillatory behavior, as the walk rarely crosses between the partitions.

The highly correlated nature of the 86-member family is due to the presence of two sublandscapes, each of which exhibits a restricted range of fitnesses. Each sublandscape also tends toward the ''partitioned'' type. Thus the walk spends most of its time in one of four fitnesses, occasionally passing between them. This landscape is visualized as a graph in Fig. 3. There are only two pathways that cross the partition, which we suggest is the cause of the high autocorrelation. These landscapes illustrate the variety that can be exhibited by minimalist models.

Table III shows how increasing sequence length allows more hydrophobic residues to be incorporated into the binding pocket. This means that longer chains are necessary in order to increase the fitness of the functional model proteins, within the currently adopted definition of fitness. Thus a drive to increase fitness will produce proteins that are longer, and so occupy a richer landscape. The new landscapes lead to more opportunity for acquiring new function after gene duplication, for example. Without this feature, scope for acquiring new function would be more limited, as our population would reach a peak of fitness on a landscape of shorter chains.

TABLE III. Normalized distribution of the sequence composition of binding pockets.

| | Number of monomers in chain | | | |
|---|---|---|---|---|
| Fitness | 22 | 23 | 24 | 25 |
| 2 | 0.028 | 0 | 0 | 0 |
| 3 | 0.044 | 0.016 | 0.031 | 0.078 |
| 4 | 0.232 | 0.200 | 0.154 | 0.245 |
| 5 | 0.480 | 0.411 | 0.357 | 0.370 |
| 6 | 0.191 | 0.274 | 0.301 | 0.199 |
| 7 | 0.025 | 0.093 | 0.143 | 0.090 |
| 8 | 0 | 0.006 | 0.012 | 0.013 |
| 9 | 0 | 0 | 0.001 | 0.003 |

TABLE IV. Largest families for each chain length.

| Number of monomers in chain | Sizes of the three largest families | | |
|---|---|---|---|
| 19 | ⋯ | 2 | 6 |
| 20 | 2 | 5 | 6 |
| 21 | 1 | 1 | 1 |
| 22 | 31 | 31 | 33 |
| 23 | 31 | 45 | 45 |
| 24 | 34 | 45 | 126 |
| 25 | 86 | 129 | 151 |

### B. Visualization of the landscape

Once all folding sequences are identified, an evolutionary landscape can be constructed by finding pairs of sequences related by single point mutations. The largest families are given in Table IV. Examples of evolutionary landscapes are shown in Figs. 3 and 4. Longer chains can support larger evolutionary landscapes. Indel mutations allow shorter chains to evolve to longer chains through a connected pathway. The increase in the size of the landscape with length would provide greater stability to mutation. This in turn would provide a driving force for evolution to greater length, even in the absence of improved function for the longer sequences. An opposing force may be a complexity catastrophe—as the sequences increase in length, they will

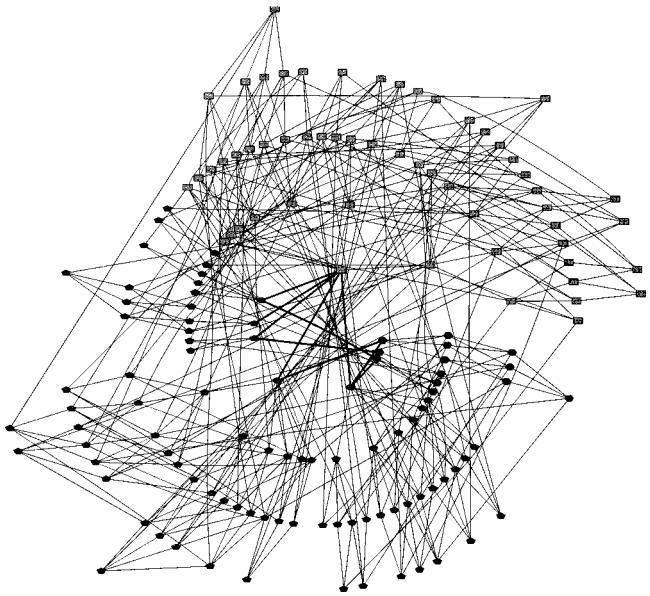

FIG. 4. 151-member family of 25-mers, drawn with the same convention as Fig. 3.

be more likely to sustain mutations and so more likely to suffer deleterious mutations. However, for the short chains that we study, an increase in length of only a small proportion of the chain leads to families which are sufficiently larger to compensate for the greater number of deleterious mutations. A similar pattern is observed for functional model proteins on the square lattice. However, the sizes of the largest landscapes on the diamond lattice grow in a smoother fashion.

We visualize the landscapes by representing a viable sequence as a node depicted according to fitness. Edges between the nodes represent allowed point mutations. We can assess various features of the landscapes by examining their topology. Some properties of our landscapes are easiest to see when considered on a qualitative level. Here we examine two landscapes. Figure 4 depicts the largest landscape of 25-mers found, with 151 members. This landscape exhibits some of the features that seem to be common to most of the larger landscapes. The landscape consists of a set of linked neutral networks. A large network of low fitness (four hydrophobes in the binding pocket) is surrounded by directly attached networks of greater fitness (five), which can then lead to a further improvement in fitness (to six). The neutral networks of highest fitness tend to be the smallest. This aspect is common to many of our landscapes—for obvious reasons, sequences that can maintain many hydrophobes in the binding pocket are rarer than those with less. Thus, the areas of highest fitness are least able to withstand the rigors of mutation. The longer chains are able to support more hydrophobes more easily (Table IV), which will lead to larger neutral networks of higher fitness. This could provide a driving force to longer chains with larger neutral networks for higher fitnesses. None of the mutations in this landscape changes the structure. This is characteristic of most landscapes on the diamond lattice. Sequences on this landscape are restricted to a single structure, which is the most designable structure found, shown in Fig. 1.

A previously discussed feature is the presence of critical edges,[26] mutations that connect two otherwise unconnected portions of the landscape. These features tend to arise on the larger landscapes. Although the landscapes on the diamond lattice were smaller than those found on the square lattice, we see an example of a similar feature in Fig. 3, where two smaller landscapes are linked through two mutations. These mutations link two different sublandscapes, each of which has its own structure and characteristics.

We observe a feature of the networks that may be considered complementary to critical pathways. The landscapes often resemble small world networks,[33] with connection topology that lies somewhere between fully regular and random. Each neutral network is connected to a number of different neutral networks by adaptive mutations of the sequences in the network. A striking feature is the presence of sequences that are able to make all these connections individually. Furthermore, these nodes often connect to each other. We refer to these nodes as ''hubs.'' Thus, if we start at a hub, we can move quickly to any of the other networks without having to cross the individual neutral networks. These features are observed in ten of the twelve families with 30 or more sequences, three or more neutral networks and sequence length 25. Even if this is just a feature of landscapes formed by the short chains we have examined, it still has implications for the evolution of the earliest proteins, in that the structure of the evolutionary landscapes naturally facilitates rapid exploration of the individual neutral networks. This will reduce the probability of becoming trapped on a local maximum, as the neutral network of global maximum fitness will be reachable by few mutations over hub sequences, even if a neutral net of lower fitness needs to be crossed.

In the case of the 151-member family, all binding pocket configurations on the landscape can be formed by making mutations within the binding pocket of a hub sequence. This sequence can sustain some mutations outside the binding pocket, however these are not necessary to support new function, and sometimes cause adaptive mutations to become lethal. It is interesting that this effect is duplicated in a landscape such as the 86-member landscape, that actually consists of two structures. In this case, after a structural mutation all the possible neutral networks of the new structure can also be reached in the same way.

### C. Structural mutations and designability

In families, previously examined on the square lattice, we frequently observed mutations that altered the structure, while maintaining or altering function. A different behavior is observed for the diamond lattice. Many landscapes are composed of a single conserved structure. For the 25-mer, the majority of landscapes consist of only one structure, with only two families (of 86 and 33 members) consisting of two structures. The difference in designability of structures may stem from the change in dimensionality, which opens up a different set of possible contact maps. The nature of contact map space is extremely important for determining the designability of a structure.

Contact map space can be imagined as the set of all structurally allowed contact maps, separated from each other by their Hamming distances. Each contact map encodes the nonbonded pairwise contacts. The Hamming distance is the number of differences between two. For example, contact maps, (1–6,3–8,6–14) and (1–6,8–14) would be three apart in contact map space, as we lose two contacts and gain one. The local nature of contact map space is critical in determining, first whether a structure can be a foldable protein structure at all, and second, whether that structure can be highly designable. Consider the nature of contact map space around a certain contact map. If there are many points nearby, that is, if there are many similar contact maps, then any sequence folded into that structure will have many alternative structures similar in energy. This means that the ground state will likely be degenerate or an energy gap will not be established, and so the structure will not be a foldable functional model protein structure.

If a contact map is sufficiently isolated, it may be encoded for by a sequence. Subsequently, if the contact map is even further isolated, then any sequences folded into this structure will be able to undergo substantial mutations before any other structures can come close in energy. This will lead

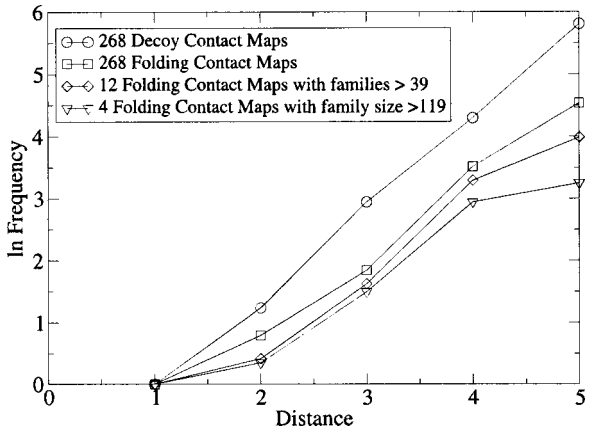


FIG. 5. Contact map space. The logarithm of the mean frequency of occurrence of structures separated from a set of structures by Hamming distance, for several sets of sequences.

to a highly designable structure. To consider this, we plot the average number of neighbors in contact map space for a variety of 25-mer structures in Fig. 5. We plot the mean frequency of contact maps that are at various Hamming distances for all 268 contact maps that are designable by at least one sequence. We compare this with a set of 268 randomly chosen, physically allowable contact maps that are not known to be encoded by a shifted-HP sequence. We constrain the set of random contact maps to be ''proteinlike'' by ensuring the number of contacts per map is distributed identically to the 268 folding structures. We also plot the mean number of neighbors for only the more designable sequences. On Fig. 5 the most designable structures are clearly more isolated from their neighbors than less designable structures. Also, the random structures have many more near neighbors than the designable structures. One implication of this is that a possible method for finding large families in sequence space would be to find the most isolated contact maps after the exhaustive enumeration of structure space. However, this process may take as long as the exhaustive enumeration of sequence space.

It may be that proteinlike features such as secondary structure and tertiary symmetries such as $\beta$-barrels, four-helix bundles and $\alpha$-helical coiled-coils emerge as a result of maintaining isolated contact maps as chain length increases. Duplication of designable substructures has been speculated to be a source of tertiary symmetries.[34] It may be that certain isolated contact maps are typical ''winning solutions.'' The duplication of the structures corresponding to these contact maps would then result in symmetries that maintain or extend the separation of a structure's contact map from surrounding maps, and so lead to the structure becoming more designable.

One aspect of designability is that very different sequences can design the same structure. This is illustrated by the 151-member family, where there are sequences which differ in nine of the 25 monomers. As noted earlier, in contrast to the diamond lattice, functional model proteins on the square lattice show much variation in structure. For example, the largest family of 23-mers has 713 members, yet the most designable structure is encoded for by only 64 sequences.

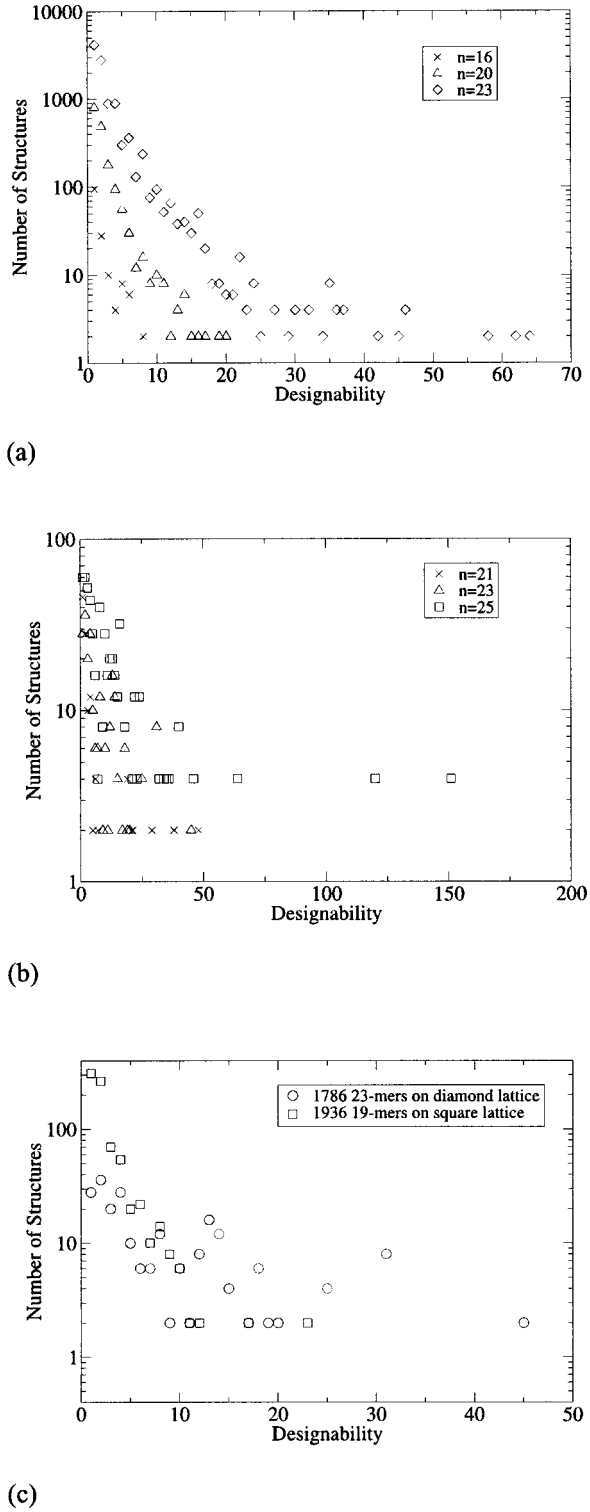


FIG. 6. The frequency of designabilities for (a) square lattice and (b) diamond lattice functional model proteins, with a comparison (c) between models on the different lattices with similar numbers of functional sequences.

Figure 6 shows the distributions of designability on the two lattices. The primary difference is that the diamond lattice demonstrates a much smaller number of structures of low designability. This corresponds to the large families where structure is highly pliable. Figure 6(c) shows that this behavior is not a simple consequence of the number of functional sequences found. The figure shows the 23-mer on the dia-

mond lattice, which yields 1786 functional sequences, and the 19-mer on the square lattice, which yields 1936. The 19-mer possesses an order of magnitude more structures of low designability than the 23-mer. It is unclear what feature of the lattices causes this behavior, or whether it is restricted to the shifted-HP model, but it raises questions for studies that rely on structural characteristics for their definitions of function. Exhaustive studies on 3D lattices are rare, with most studies concentrating on maximally compact structures on a cubic lattice.[20–22] We suspect that for the square lattice there are a large number of structures that are somewhat isolated in contact map space, and so are foldable, but poorly designable. On the diamond lattice, this may be different; the structures may be either highly isolated or well surrounded, with intermediate structures more of a rarity.

## IV. CONCLUSIONS

In this study, we have extended our previous work[25,26] by the study of functional model proteins on a 3D lattice. We have mutated shorter sequences in order to examine families of longer sequences. The requirement for function is one which makes functional model proteins rarer in our evolutionary landscapes than they would otherwise be. However, we believe it leads to a more accurate characterization of evolutionary landscapes than models that consider that all folding proteins are functional.

The difference between the diamond lattice and the square lattice with respect to designability was highlighted. Many previous studies represent function in terms of preservation of structure. Perhaps conclusions drawn from these studies should be reconsidered in light of the evidence that the degree of structural mutations is tightly coupled to the choice of lattice.

Designable structures were characterized as those which have highly unusual contact maps. Crippen[35] applied an empirical function to predict the number of structures with a given contact map on the cubic lattice. Examples of large deviations from this function were found, i.e., contact maps which were much more or less common than anticipated. The complexity of contact map space is important, as the evolutionary landscapes of minimalist models are strongly dependent on its nature.

Several conclusions drawn from our square lattice studies remain robust with respect to the choice of lattice. We observe that the size of our landscapes grows with chain length, with the landscapes becoming less rugged. Longer chains are able to support more functional diversity, incorporating more hydrophobic character into the binding pocket. Longer chains also demonstrate new features, with 24-mers supporting closed pockets (those surrounded by four or more monomers), and 25-mers also supporting two binding pockets. These characteristics of longer chains can provide impetus for evolution to grow the chain, which will provide access to larger, more structurally diverse landscapes, which will in turn allow proteins to acquire new function.

The previously observed frequency of critical pathways[26] may have been a feature of the square lattice, due to its propensity for structures of low designability, which may link up otherwise separated families by forming intermediate structures. However, similar features are still observed, for example with the 86-member family of 25-mers (Fig. 4). This family consists of two structures organized into two subfamilies connected by two edges. This is qualitatively similar to the effect of known mutations of Arc protein.[27] A mutant of Arc where a hydrophobe is swapped with an adjacent polar residue (Asn11Leu and Leu12Asn mutations) has a different tertiary structure. A sequence with just the Asn11Leu mutation is stable in either structure under different conditions. While the simulation of such characteristics is currently beyond the scope of our investigation, the underlying structures of the landscapes are able to reflect some aspects of real landscapes.

## ACKNOWLEDGMENTS

We thank the BBSRC Bioinformatics Committee for a Ph.D. studentship and Dr. Natalio Krasnogor for useful discussions.

[1] M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1983).
[2] S. Wright, Proceedings of the Sixth International Congress on Genetics, 1932, p. 354.
[3] S. A. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford University Press, Oxford, 1993).
[4] T. Yomo, S. Saito, and M. Sasai, Nat. Struct. Biol. **6**, 743 (1999).
[5] N. Go, Int. J. Pept. Protein Res. **7**, 313 (1975).
[6] K. A. Dill, S. Bromberg, K. S. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, Protein Sci. **4**, 561 (1995).
[7] G. Tiana, R. A. Broglia, and D. Provasi, Phys. Rev. E **64**, 011904 (2001).
[8] P. D. Williams, D. D. Pollock, and R. A. Goldstein, J. Mol. Graphics Modell. **19**, 150 (2001).
[9] D. M. Taverna and R. A. Goldstein, J. Mol. Biol. **315**, 479 (2002).
[10] M. Eigen, Naturwissenschaften **10**, 465 (1971).
[11] H. S. Chan and K. A. Dill, Proteins: Struct., Funct., Genet. **24**, 335 (1996).
[12] K. A. Dill, Biochemistry **29**, 7133 (1990).
[13] W. Kauzmann, Adv. Protein Chem. **14**, 1 (1959).
[14] D. A. Moffet and M. H. Hecht, Chem. Rev. **101**, 3191 (2001).
[15] S. A. Marshall and S. L. Mayo, J. Mol. Biol. **305**, 619 (2001).
[16] H. J. Cordes, N. P. Walsh, C. J. McKnight, and R. T. Sauer, Science **284**, 325 (1999).
[17] Y. Mandel-Gutfreund and L. M. Gregoret, J. Mol. Biol. **323**, 453 (2002).
[18] E. Bornberg-Bauer, Z. Phys. Chem. (Munich) **216**, 139 (2002).
[19] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[20] M. H. Hao and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **93**, 4984 (1996).
[21] E. I. Shakhnovich and A. Gutin, J. Chem. Phys. **93**, 5967 (1990).
[22] E. I. Shakhnovich, Folding Des. **3**, R45 (1998).
[23] A. Irbäck and C. Troein, J. Biol. Phys. **28**, 1 (2002).
[24] K. Yue, K. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **92**, 325 (1995).
[25] J. D. Hirst, Protein Eng. **12**, 721 (1999).
[26] B. P. Blackburne and J. D. Hirst, J. Chem. Phys. **115**, 1935 (2001).
[27] F. T. Wall and F. T. Hioe, J. Chem. Phys. **74**, 4410 (1970).
[28] N. E. G. Buchler and R. A. Goldstein, J. Chem. Phys. **111**, 6599 (1999).
[29] H. Kaya and H. S. Chan, J. Mol. Biol. **315**, 899 (2002).
[30] L. Li, L. A. Mirny, and E. I. Shakhnovich, Nat. Struct. Biol. **7**, 336 (2000).
[31] M.-H. Hao and H. A. Scheraga, J. Mol. Biol. **227**, 973 (1998).
[32] J. J. Chou and E. I. Shakhnovich, J. Phys. Chem. B **103**, 2535 (1999).
[33] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[34] T. Wang, J. Miller, N. S. Wingreen, C. Tang, and K. A. Dill, J. Chem. Phys. **113**, 8329 (2000).
[35] G. M. Crippen, J. Chem. Phys. **112**, 11065 (2000).