Evolution of functional model proteins

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

(Received 13 December 2000; accepted 9 May 2001)

The distinct influences of function, folding, and structure on the evolution of minimalist model proteins are studied by characterization of their evolutionary landscapes. Chains of up to 23 monomers on a two-dimensional square lattice are investigated by exhaustive enumeration of conformation and sequence space. In addition to common aspects of minimalist models, such as unique, stable native states and cooperative folding, functional model proteins have the novel feature of an explicit binding pocket. Fitness is defined through simple, physical characterization of the binding pocket. We characterize various properties of functional model proteins, focusing on their evolutionary landscapes, as defined by single point mutations, insertions, and deletions. The longer chains more closely resemble real proteins, having richer functional diversity and forming larger families of sequences. Although regions of evolutionary landscapes are often highly interconnected, we also observe so-called critical pathways, where evolution can only proceed through a single set of mutants. © *2001 American Institute of Physics.* [DOI: 10.1063/1.1383051]

I. INTRODUCTION

Molecular recognition, folding, and evolution are fundamental phenomena associated with proteins. Highly simplified, so-called minimalist, models of proteins are emerging as a means of studying these three intricately linked features of proteins. Minimalist models entail simplified representations of protein sequence and structure, with a reduced alphabet of amino acids and the restriction of the protein chain so that it lies on points of a lattice. Minimalist models have been most widely applied to studying the protein folding problem, as described in a recent review (Ref. 1, and references therein). The review highlights how the interplay between experimental studies and theoretical ideas that have emerged from the study of minimalist models has led to a deeper understanding of the folding process.

A more recent application of minimalist models has been the investigation of aspects of ligand binding.[2] Various types of binding behaviors have been explored using a modification of the two-dimensional HP lattice model. The standard model of hydrophobic (H) and polar (P) residues[3] was extended to include a third, ligand (L), monomer type.[2] One application of the model was the investigation of aspects of induced fit. Similar issues have been investigated using an adaptation of the folding funnel concept, to provide a framework for understanding ligand binding and binding mechanisms.[4]

Evolutionary aspects of proteins are increasingly investigated using lattice models.[5–7] Such studies tend to use definitions of fitness or function that relate to properties, such as structural integrity or characteristics of the folding process. The neutral theory of evolution[8] is now widely accepted, although not universally, and plays a key role in the understanding of many evolutionary questions. This theory proposes that not every mutation is adaptive, changing fitness, but that many mutations are neutral, conserving fitness. Evolution is suggested to proceed through neutral mutations, where adaptive mutations are not possible. Understanding the balance between neutral and adaptive mutations is an area of active interest. One recent study[5] explored the sequence landscape of a 48-residue, 20-letter alphabet, cubic lattice model. Neutral mutations were identified as those that preserved foldability into the same structure. A variety of evolutionary simulations have been used to study the distribution of structures in the evolution of a 25-residue, maximally compact protein model confined to a two-dimensional square lattice.[6] The fitness of sequences was taken to be related to the ability of the protein to fold. In another characterization of evolutionary landscapes,[7] 18-mer HP and AB models on a two-dimensional square lattice were studied and several definitions of fitness were explored including a step function. A neutral mutation was defined as one that encoded for the same ground state structure.

From these investigations and other earlier work that we have briefly discussed previously,[9] it is clear that studies of minimalist models of proteins are contributing to our understanding of the nature of evolutionary landscapes of proteins. The simplicity of the models permits a large number of sequences to be studied; this is an important issue for the development of a general framework for protein evolution. Aspects of folding, function, and evolution were brought together in an earlier study,[9] in which we introduced functional model proteins. Functional model proteins are not maximally compact and contain an unoccupied lattice site at least partially surrounded by the rest of the protein chain. This provides a binding pocket. The presence of a binding pocket is required for a protein to be deemed functional. Other more common criteria, discussed later, must also be met for the protein to viable. These include the requirement for a unique, nondegenerate lowest energy ground state. To impose cooperative folding, we require an energy gap between the native state and the first excited state. Thus, func-

[a)] Electronic mail: jonathan.hirst@nottingham.ac.uk tional model proteins allow us to investigate evolutionary landscapes with respect to function as a distinct feature from structure or foldability. Quantification of the hydrophobic character of the binding pocket allows us to define a simple, physical scale of fitness and thus move beyond two-state (fit/not-fit) step-function models of fitness.

In the present study we do not address the kinetic accessibility of the binding pockets. This is a dynamical issue related to fluctuations of the chain, which would require extensive calculations beyond the scope of the current investigation. We limit ourselves to thermodynamic aspects of binding. Many cavities in real proteins are able to open and act as binding pockets, and so our definition of binding pockets is a reasonable point from which to examine evolutionary issues. However, we do include an analysis of model proteins with pockets located on the surface and directly accessible, to compare and contrast the properties of open and closed pockets.

One of the virtues of minimalist models of proteins is that the different aspects of the models can be explored in detail and in a controlled fashion. Thus, the robustness or generality of conclusions about evolutionary landscapes may be investigated with respect to the type of amino acid alphabet, the length of the protein chain, the definition of fitness, and the type of lattice. In this sense, many of the studies of the evolution of minimalist models of proteins are complementary, contributing to a consensus understanding of the most general characteristics of the evolutionary landscapes of real proteins. Hence, in this study, we explore the nature of the evolutionary landscape of functional model proteins. We investigate how an explicit definition of function and a multiple-valued scale of fitness affect the nature of the evolutionary landscape. This is one novel aspect of the work and adds the realism of explicit function to these models of proteins, albeit in a highly simplified manner. We have previously studied chains of up to length 20.[9] Here, we present a more detailed characterization and introduce some algorithmic improvements to increase the length of the chains that we investigate to 23. This in turn increases the sizes of the observed protein families, again, in a modest way, more closely mimicking real proteins.

II. METHODS

A. Minimalist models

Various amino acid alphabets may be used in minimalist models of proteins.[10] One of the simplest, the HP model,[3] derives from the insight that the hydrophobic interaction is a major determinant of protein folding.[11,12] This view has found widespread theoretical and experimental support, and continues to be useful in guiding experimental studies.[13] In the HP model, the interaction energy between two hydrophobic (H) residues is $-|\epsilon|$, or -1, after scaling. The energies of all other possible interactions, those involving at least one polar (P) residue, are zero. Thus, the HP model contains attractive and neutral interactions, which tends to produce maximally compact lowest-energy conformations. A model of proteins with binding pockets, empty lattice sites surrounded by the rest of protein chain, requires the introduction of repulsive interactions.

We employ a shifted-HP model,[10] so-called because the average interaction energy is shifted from 1/3 in the standard HP model to zero, through the introduction of repulsive interactions. As discussed elsewhere,[9,10] this leads to an interaction energy matrix of the form:

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

The interaction energy between two H residues, $E_{\mathrm{HH}}=-2$. For all other interactions $E_{\mathrm{HP}}=E_{\mathrm{PH}}=E_{\mathrm{PP}}=+1$.

In minimalist models, the protein chain is restricted to lie on a lattice. This discretized model permits the exhaustive enumeration of all possible conformations. We employ the two-dimensional square lattice. As noted by Dill,[14] these models are simplified in their representation of atomic details and energies, but refined in that their full conformational space and full sequence space can be explored exhaustively without sampling or approximation. Full conformational enumeration is important in the study of functional model proteins, which are not maximally compact. Full enumeration of sequence space is advantageous, as we seek to characterize the nature of evolutionary landscapes.

Conformations are enumerated by generating all possible self-avoiding random walks on the square lattice. The algorithm used has been described previously.[9] A contact map, a list of nearest-neighbor nonbonded monomers, is constructed from each walk. It is not necessary to evaluate the energy for every conformation,[10] merely for each unique contact map. The degeneracy of each contact map is recorded, in order to determine the degeneracy of the lowest-energy conformation.

B. Computational strategies

The next step is an exhaustive search through sequence space, whereby the energy of each possible sequence is computed for each possible conformation. Several observations can be exploited to improve the computational efficiency of the exhaustive search. As mentioned, we consider only unique contact maps rather than individual conformations. Another gain in efficiency comes from only evaluating one of each pair of symmetrical sequences. Symmetric pairs of sequences adopt symmetry-related conformations, with identical energies. If one of the pair is a functional model protein, then so is the other. In our implementation, sequences are classified as left handed, right handed, or symmetric, based on the location of the majority of H residues. So, HHPPPH is classified as left handed, and HPPPHH is right handed. In the enumeration of sequence space only left-handed and symmetric sequences are considered. This results in almost a factor of 2 in efficiency, as the overhead from classification is minimal, and only very few sequences are symmetrical.

The density of states of each sequence is determined. Following current thinking on the nature of proteins, we define a functional model protein based on a number of criteria. Whilst these criteria might be the subject of ongoing debate,

TABLE I. Sequence composition and compactness of 21-mer functional model proteins.

No. Contacts	4	5	6	7	8	9	10	11
No. H residues								
5	0	0	0	0	2	0	0	0
6	0	0	6	0	6	0	0	0
7	0	16	25	6	38	10	0	0
8	0	10	37	44	82	118	24	0
9	0	0	21	182	196	352	72	0
10	0	0	0	162	342	700	108	16
11	0	0	0	38	258	1370	178	24
12	0	0	0	2	30	1304	254	6
13	0	0	0	0	4	580	266	8
14	0	0	0	0	0	76	108	24
15	0	0	0	0	0	0	2	6

TABLE II. Number of conformations, Ω, number of contact maps, U, and number of singly degenerate contact maps, $U_{[g=1]}$ as a function of chain length, n.

N	Ω	U	$U_{[g=1]}$	No. functional model proteins
19	15 582 342	363 010	140 708	1 936
20	41 889 578	910 972	400 152	3 953
21	112 212 146	1 953 847	742 238	7 113
22	301 100 754	4 868 343	2 068 843	10 605
23	805 570 061	10 513 772	3 907 514	31 146

they are a reasonable basis for a minimalist model of proteins. A functional model protein is required to have a nondegenerate ground state. This criterion precludes the possibility of conformationally diverse loop regions, but is a good starting point for a simple model. To ensure cooperative folding, functional model proteins must have an energy gap between the native state and the ensemble of non-native states. Cooperative folding is a key feature of the thermodynamic behavior of real proteins. We chose one of several possible simple criteria to model cooperativity, which if not universally accepted, does have some support.[15] Finally, the presence of a binding pocket is required, to confer function on our functional model proteins, and so permitting the definition of fitness and the subsequent analysis of the fitness or evolutionary landscapes of functional model proteins.

In an attempt to attenuate the exponential growth of the computational problem, we have explored several strategies. The first, and most successful, method is the exclusion of symmetry-related sequences, as already described. We have investigated the potential of a strategy based on the composition of a protein sequence. Intuitively, a sequence that is entirely hydrophobic or entirely polar will not be a viable protein. Chains that contain almost exclusively H or P residues also are not viable. As has been observed previously,[16,17] increasing the number of H residues in a sequence tends to reduce its stability; alternative conformations with low energies appear. The composition of 19-mer, 20-mer, and 21-mer sequences was analyzed. The scope for exclusion from the enumeration of sequences on the basis of sequence composition was assessed. We also examined compactness as a possible means of excluding some conformations or contact maps from the exhaustive enumeration.

The most effective strategy found was a combined approach based on sequence composition and conformational compactness. As the number of H residues increases in a sequence, the number of contacts tends to increase, as there are more potential favorable HH contacts. Thus, the sequence composition and compactness strategies can be profitably combined. A greater number of noncompact conformations can be ignored in the calculation of the ground and first excited states of hydrophobic-rich sequences. The number of contacts is related to the compactness of the chain. This is more satisfactory as a measure of compactness than the radius of gyration, R_G, which requires time to calculate and is not directly related to the number of potential HH contacts. The variation of the number of contacts with sequence composition was analyzed for 19-mers, 20-mers, and 21-mers. Data for the latter are presented in Table I, which shows the frequency of occurrence of functional model proteins with respect to the number of contacts and sequence composition. The shading in Table I shows areas, with a margin for variation, which may be excluded from the enumeration procedure. There are no functional model proteins with many contacts and few H residues, or conversely, few contacts and many H residues. For the 21-mer, the combined strategy reduced the computational cost of enumeration by about 5%. Enumeration of 22-mers and 23-mers was performed using just the combined strategy. The enumeration of 23-mers took about four weeks using four Compaq Alpha ES40 processors.

C. Function, fitness, and evolution

We adopt a simple model of function, based on nonspecific hydrophobic binding. The efficacy of function, or the fitness, of a functional model protein is directly proportional to the number of H residues in the binding pocket, and may vary between 0 and 8. The corners of the binding pocket are included in this definition to expand the range of fitness. Chains of moderate length can sometimes accommodate two or more binding pockets.[9] In this case, the most hydrophobic pocket is used to compute the fitness of the functional model protein. Pockets may be characterized as either surface or cavity, depending on their position. Cavity pockets are completely surrounded by the chain in the native state, surface pockets are only surrounded by the protein chain on three sides. The majority of our investigation includes both cavity and surface pockets. However, we also present a separate

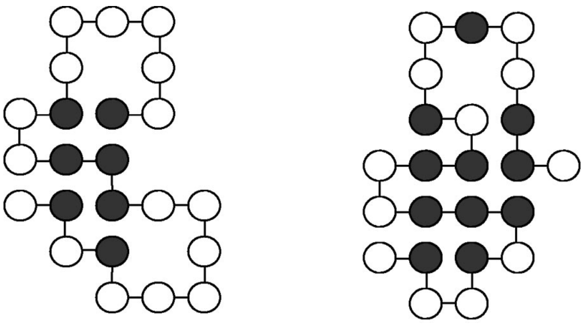

FIG. 1. Illustrative examples of 23-mer functional model proteins. Black monomers are hydrophobic, white monomers are polar.

TABLE III. Normalized distribution of sequence composition of binding pockets. Fraction of sequences with n_H H residues in the binding pocket.

	Chain length				
n_H	19	20	21	22	23
0	0.001	0.002	0	0.002	0
1	0.010	0.012	0.008	0.013	0.012
2	0.152	0.163	0.091	0.126	0.095
3	0.515	0.565	0.497	0.555	0.413
4	0.201	0.166	0.201	0.157	0.229
5	0.121	0.090	0.197	0.144	0.235
6	0	0.002	0.006	0.003	0.013
7	0	0	0	0	0.003

analysis of open pockets. Other definitions of fitness could have been used, and may be worth exploring in due course. The model adopted in this study reflects the spirit of minimalist models, in that it is simple yet based on a physical principle.

Once all functional model proteins of a given chain length were identified, through application of the criteria regarding stability, folding, and function, the evolutionary landscape was characterized. Pairs of functional model proteins related by a single point mutation were identified. These were used to construct families of proteins. The distribution of the sizes of families was analyzed. Another interesting property is the interconnectedness of protein families. A more interconnected family may be expected to exhibit more facile evolution of new structure or function. Families were represented as graphs, and the interconnectedness was computed as the mean number of edges (single point mutations) possessed by a node (sequence) on the graph. The stability of the function of a protein with respect to mutation was assessed by examination of the proportion of allowed mutations that maintained function. We analyze the ratio of neutral to adaptive mutations within families and across the evolutionary landscape. In addition, we compare the evolutionary landscapes of all functional model proteins with those for proteins with surface pockets only (i.e., excluding cavities).

TABLE V. Characterization of adaption in terms of structure or function expressed as the fraction of nonlethal single point substitutions.

n	Neutral (with respect to structure and function)	Adapt structure only	Adapt function only	Adapt structure and function
20	0.66	0.07	0.19	0.08
21	0.67	0.05	0.19	0.08
22	0.70	0.08	0.13	0.09
23	0.59	0.14	0.18	0.09

III. RESULTS

A. Sequence and structural characterization

Our results and discussion focus primarily on chains of length 19–23. Functional model proteins of lengths 11–20 have been studied previously.[9] We present some new, more detailed analyses of the 19-mer and 20-mers, in addition to the data on the longer chains. In Table II, we show how the conformational space grows with chain length. The number of different contact maps grows exponentially, but not as rapidly as the number of conformations. The requirement of a unique ground state limits the number of possible native structures to the nondegenerate contact maps, but in order to identify the lowest energy conformation we must also consider the degenerate contact maps.

The sequence compositions of the viable functional model proteins of lengths 19–21 were analyzed. The 21-mers have the broadest range in composition, containing between 5 and 15 H residues. The sequence compositions are a little narrower than the binomial distributions for the entire sequence spaces. For the entire sequence space of 21-mers, only ~5% of sequences have less than 5 H residues or more than 15. Thus, restricting the enumeration of sequences based on composition alone offers little gain in computational efficiency, if one wishes to identify all functional model proteins.

The number of functional model proteins grows exponentially with chain length (Table II), and the fraction of all possible sequences that are actually viable is approximately constant. Two examples of 23-mer functional model proteins are illustrated in Fig. 1. It is evident that functional model

TABLE IV. Characteristics of evolutionary landscapes.

	No. functional model proteins		Mean allowed mutations/sequence		Neutral:adaptive mutations		Mean number of binding pockets	
n	Total	Largest family	All seqs	Largest family	All seqs	Largest family	All seqs	Largest family
16	289	18	1.42	1.28	1.34	0.53	1.05	1.67
17	652	30	1.90	1.70	2.15	0.82	1.06	1.13
18	819	14	1.28	2.29	5.92	0.78	1.05	1.14
19	1 936	59	1.86	3.29	2.61	1.49	1.13	1.14
20	3 953	576	1.74	3.19	2.77	1.72	1.12	1.16
21	7 113	71	1.87	3.43	2.81	1.77	1.14	1.14
22	10 605	80	1.41	3.05	3.43	1.71	1.17	1.33
23	31 146	713	2.39	3.98	2.73	2.51	1.21	1.21

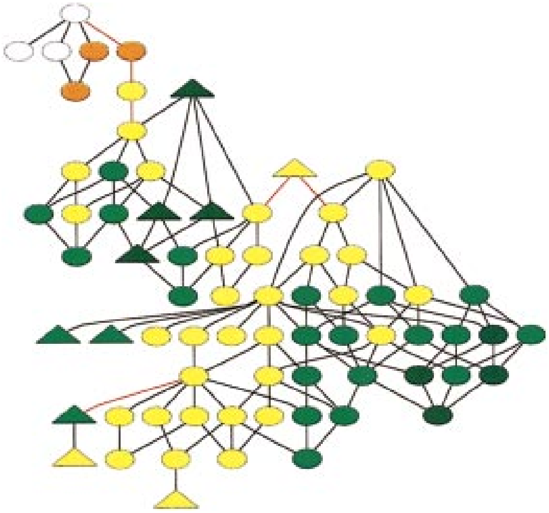

FIG. 2. (Color) 71-member family of 21-mers. Nodes correspond to functional model proteins; edges correspond to single point mutations connecting two viable sequences. Nodes are color coded based on function (the number of H residues in the binding pocket): white=1, orange=2, yellow=3, light green=4, dark green=5, red=6. The shape of a node indicates the number of binding pockets: ellipse=1, triangle=2, rectangle=3. Red edges are described in the main body of the paper.

proteins can be quite diverse. One example shows three binding pockets, two completely enclosed by the rest of the chain and one open pocket, surrounded on three sides. Functional model proteins are not maximally compact, but tend to have hydrophobic cores, which provide stability. Although there may be a weak correlation between chain length and hydrophobic content, our results are broadly in agreement with the observation that the fraction of hydrophobes does not grow significantly with chain length.[18]

B. Function and evolutionary characterization

The data in Table III present an analysis of the functional diversity of functional model proteins, as measured by the number of H residues in the pocket, mimicking a simple physical model of nonspecific hydrophobic binding. Functional model proteins favor binding pockets containing at least two H residues, to facilitate packing of the binding pocket or loop to the rest of the protein. Three H residues in the binding pocket is consistently the most common sequence composition. Longer chains seem able to support more hydrophobic pockets, leading to a modest increase in functional diversity. This perhaps also leads to a selective pressure to increase chain length, as our basic definition of function means that functional model proteins with pockets of greater hydrophobic character are fitter.

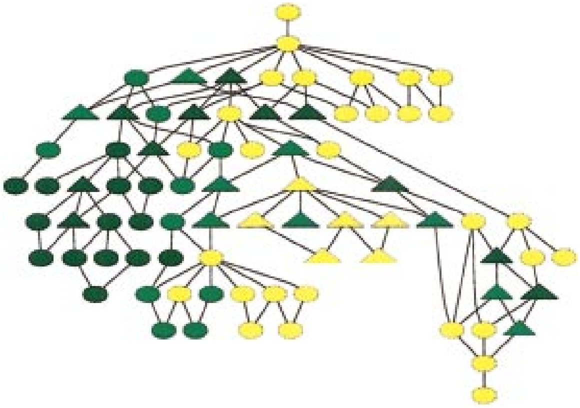

FIG. 3. (Color) 80-member family of 22-mers, drawn with the same convention as Fig. 2.

The distribution of sizes of families of proteins is one characteristic that determines the nature of the evolutionary landscape. Proteins in large families have a greater potential for evolution and are more robust with respect to single point mutations. Several properties of the evolutionary landscapes, including the size of the largest family, are given in Table IV. The table includes some new data on the shorter chains, and some data that have been corrected from our earlier study.[9] The sizes of the largest families increase with chain length. Most families have symmetry-related analogs whose members are the symmetry-related sequences. Occasionally, a palindromic sequence connects a pair of symmetry-related families, forming a single larger family. This happens in the case of the 576-member family of 20-mer, partly explaining its unusual size. The results for the 21-mer and 22-mer follow the underlying trend of a more modest increase in the size of large families with chain length, underscoring the value of studying longer chains and extending our previous study[9] beyond 20-mers. The 23-mer produces many large

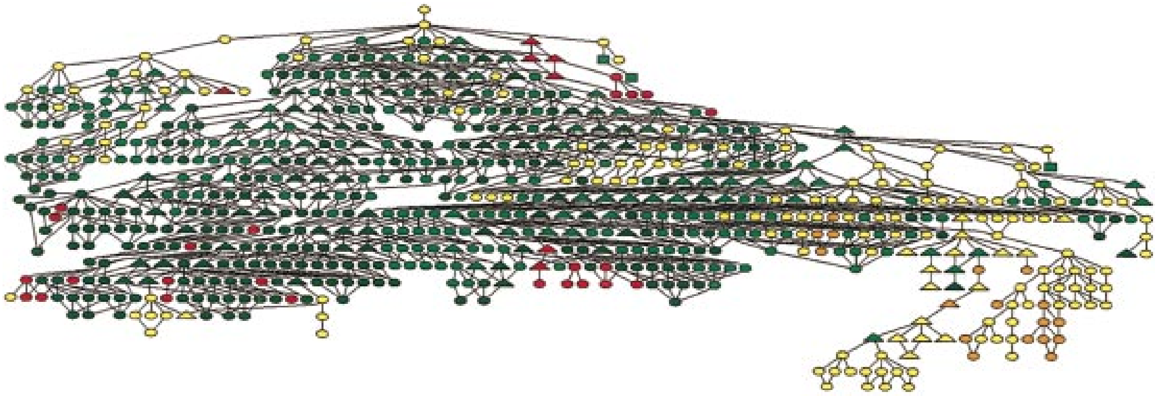

FIG. 4. (Color) 713-member family of 23-mers, drawn with the same convention as Fig. 3.

families. There are 26 families of 23-mers with more than 100 members each.

Another interesting measure is the interconnectedness of the families. Interconnectedness, like size, confers evolutionary flexibility and robustness. If a protein family is represented as a graph, with the sequences as nodes and single point mutations as edges, interconnectedness may be simply characterized by the mean number of edges connected to a node, or the mean number of nonlethal mutations per sequence. The 23-mers show the most interconnected evolutionary landscape. To characterize evolutionary landscapes more fully, one needs to consider not only the numbers of nonlethal mutations, but also their nature.

Functional model proteins, with their explicit definition of function, allow the characterization of mutations as either neutral (conserving function) or adaptive (changing function) in a way that connects more directly with the concepts of function and fitness than measures related to structure preservation or foldability. Table V demonstrates that in practice there is a clear distinction in our model between mutations which affect structure and those which affect function.

Table IV shows that there is a slight tendency for the ratio of neutral:adaptive mutations to increase with chain length across entire evolutionary landscapes. This trend is much more pronounced for the largest families, indicating a greater tolerance to mutation and leading to a greater clustering of function within the families. In terms of the local fitness landscape, this corresponds to the landscape becoming less rugged. This is despite the greater diversity of function in the longer proteins. One explanation for this may be that for longer chains a smaller proportion of the protein defines the binding site, and thus a greater proportion of nonlethal mutations will fall outside this region and may have no effect on the binding site. This suggests that longer, more sophisticated, more realistic models of proteins might be expected to exhibit a still larger proportion of neutral: adaptive mutations, supporting the idea that random drift along neutral variants is important, as asserted by the neutral theory of evolution.[8]

Graphs of the largest families of 21-mers, 22-mers, and 23-mers are depicted in Figs. 2, 3, and 4. Some landscapes show what we term ''critical edges''—edges that bridge two otherwise unconnected areas of the landscape. Edges connecting a single sequence to the landscape are excluded from this definition. The critical edges are marked in red in Fig. 2. Several consecutive critical edges form a critical pathway, which sometimes connects two different neutral networks, thus controlling evolution between these networks and the acquisition of new function or improved fitness. One significance of critical pathways could be in the prevention of drug resistance. If resistance to a drug evolves through a critical pathway, the several mutants along the pathway could be specifically targeted, thereby reducing the likelihood that drug resistance would evolve.

For the 22-mer there are 30 families with 20 or more members, and of these families 20 have at least one critical edge and 7 have at least one pathway of three or more consecutive critical edges. Of the 166 families with 20 or more members found for the 23-mer, 108 have at least one critical

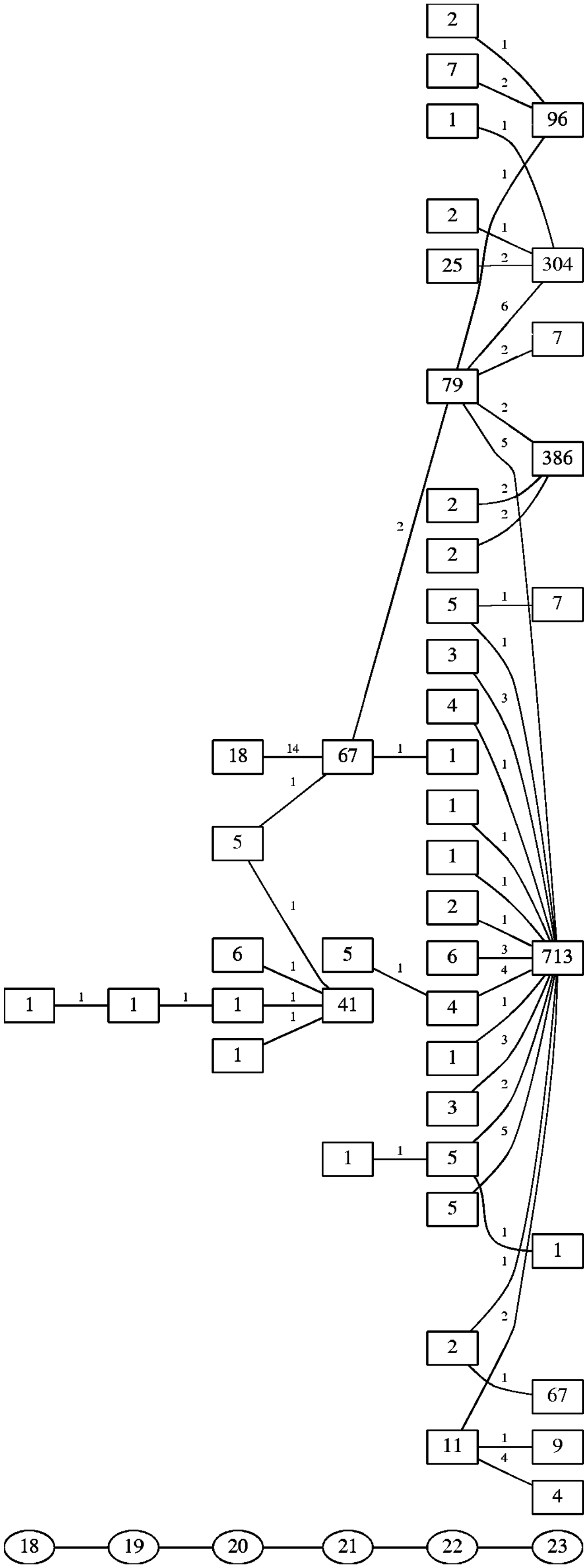

FIG. 5. 1915-member family. The elliptical nodes show the chain length. Each rectangular node is a subfamily, labeled with the number of members. Each edge indicates one or more possible insertions/deletions, and is labeled with the number of connections between each family.

TABLE VI. Characteristics of evolutionary landscapes of surface pocket proteins.

	No. functional model proteins		Mean allowed mutations/sequence		Neutral:adaptive mutations		Mean number of binding pockets	
n	Total	Largest family	All seqs	Largest family	All seqs	Largest family	All seqs	Largest family
19	340	39	1.98	2.92	2.07	1.59	1.01	1
20	445	17	1.34	2.47	3.52	9.5	1.01	1
21	1 842	54	1.91	2.89	3.66	4.57	1.05	1
22	2 592	56	1.47	3.5	5.24	4.44	1.06	1
23	7 464	105	1.81	3.94	3.87	4.45	1.08	1

edge, and 28 have at least one pathway of three or more edges. Critical pathways of seven and eight edges exist for the 22-mer and 23-mer. Thus critical edges and pathways are a relatively common feature of the landscapes, and could be visualized as restricting walks up a ''peak'' on rugged evolutionary landscapes to certain paths where fitness drops off either side. Interestingly, as we increase chain length not only are the families more highly interconnected, but we also see an increase in the number of critical pathways.

We have extended our evolutionary analysis beyond single point mutations, to include insertions and deletions. Insertion is clearly an important mechanism for increasing the length of the earliest proteins. When insertions and deletions are allowed, the effect is to link up the families already identified. The largest range of sequence lengths spanned by insertions and deletions is 18–23 (Fig. 5). This family has 1915 members and is the largest found. For this particular evolutionary pathway there is a required deletion from a family of 21-mers to a family of 20-mers before evolution to longer chains can proceed. In total, 11 families with sizes greater than 200 were observed. The inclusion of indels clearly makes the evolutionary landscape less fragmented. Further investigation into the effects of indels in our model is currently underway.

C. Characterization of surface pockets

We include an analysis of just the surface binding sites for chains of length 19–23. This is a subset of the sequences considered so far, and a corresponding decrease in family sizes is observed (Table VI). The functional diversity is also reduced, due to the open pockets being defined by between five and seven residues surrounding the binding pocket rather than the eight in closed pockets. However general trends are conserved. The mean allowed mutations per viable sequence closely follows the pattern already seen (Table VI). As expected, the distribution of the number of hydrophobes in the binding site is skewed toward lower numbers, but is still qualitatively similar. One interesting difference is that proteins with surface pockets display a greater clustering of function (Fig. 6). This is also shown in the elevated ratio of neutral:adaptive mutations. An explanation for this could be that because surface binding pockets are by definition small and at the edge of the protein, they are more likely to be formed by a single section of the chain rather than the coalescence of two or more remote sections. In this respect the cavities more closely resemble real protein active sites which often consist of remote sections of the chain brought together. This means that structural mutations are less likely to occur in a region of the chain near the binding pocket and so change the function. Indeed, approximately 2% of viable mutations adapt both structure and function for surface pockets, compared with approximately 8% when cavities are also considered.

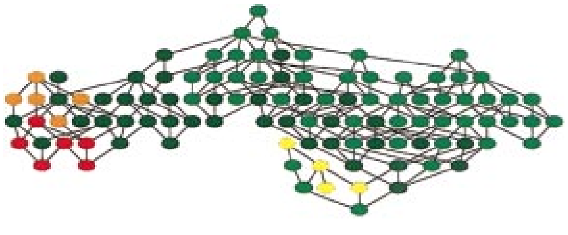

FIG. 6. (Color) 105-member family of 23-mers with surface pockets, drawn with the same convention as Fig. 3.

IV. CONCLUSIONS

In this study, we have characterized properties of functional model proteins. We have examined their sequence composition and structural features, and exploited some of them to speed up searches of conformation and sequence spaces. The explicit definition of function, as distinct from structure or other properties, is a novel aspect of functional model proteins. This provides an important additional element to minimalist models of proteins, allowing us to explore separately functional and structural diversity, and providing a finer scale of fitness to the resulting evolutionary landscapes. The importance of evaluating function as well as structure is borne out by the observation that mutations increasing stability may decrease or even abolish the activity of a protein.[19]

The potential for a detailed connection between lattice-based studies and real proteins continues to grow with the wealth of emerging experimental data, from genomic projects and combinatorial approaches. Comparisons of evolutionary data from lattice studies and real proteins often find the two in qualitative agreement. The significance of non-native interactions in the folding nucleus during evolution has been evaluated both experimentally and with a lattice model with qualitative agreement seen between the two.[20] Analysis of analogous proteins using lattice models reveals a bimodal distribution of frequency of conservation at core sites,[5] which is also observed in an analysis of structures taken from the Protein Data Bank.[21] A related study on lattice models focusing on the importance of topology on the nature of kinetically important residues in the folding nucleus found qualitatively similar features in real proteins.[22]

The use of reduced alphabets in evolutionary studies is also supported by various investigations. Babajide *et al.*[23] used a statistical measure of nativelike folding for a range of sequences against structures taken from the Protein Data Bank. They found evidence for large evolutionary landscapes in sequence space, even when reduced alphabets of hydrophobic and polar residues were considered. Of a library of binary patterned (polar and nonpolar) proteins,[13] half demonstrated cooperative thermal denaturation, some of which were also fully monomeric in solution, leading to the conclusion that they display nativelike folding.

Bornberg-Bauer and Chan[7] have investigated the nature of the evolutionary landscapes of the HP model and AB model. The latter, in which A–A and B–B interactions are favorable, exhibits a more fragmented landscape. The AB model provides a useful reference point, although it is widely regarded as a poor model for real proteins. The evolutionary landscapes we have observed are not as fragmented as those seen for the AB model. A detailed study of two letter alphabets[24] suggests that the nature of the evolutionary landscape of shifted HP models lies between the extremes of the HP and AB landscapes.

We have exhaustively enumerated chains up to length 23, somewhat longer than typical studies focused on the evolutionary context. We find that new trends emerge with the longer chains. The growth of the size of the largest family is not quite as rapid as suggested by extrapolation from chains up to length 20. The size of neutral networks grows with chain length, as does the functional diversity. Longer chains are more realistic models of proteins and lead to larger families with properties closer to real proteins. Much longer chains have been studied in other contexts, usually related to aspects of folding, using sampling methods, such as Monte Carlo, or heuristic approaches.[25,26] Clearly, such approaches warrant investigation. In the same vein, other amino acid alphabets, other lattices, and other criteria defining a functional model protein should also be explored.

In light of these issues our study has adopted a reasonable starting model, incorporating unarguably important aspects of proteins, although how these features are best modeled remains open to debate. The model shows the fitness landscape becoming less rugged with increasing chain length, and in the large families, more nonlethal mutations are available. Thus, with increasing chain length our model acquires a greater propensity for neutral mutations rather than adaptive or lethal mutations. At longer chain lengths, it is easier to acquire more than one binding pocket, and a greater diversity of function is seen. This suggests that selective pressure might drive proteins to longer chains, which in turn are more stable to mutation and have a greater opportunity for adaption.

The presence of ‘‘critical pathways’’ indicates that there may be times where evolving populations of proteins are restricted to a few sequences through which they must evolve in order to acquire new or improved function. In the language of fitness landscapes,[27] the length of the pathway would be a factor in determining whether a gene can traverse it. If the pathway is too long, then perhaps too many of the proteins in the evolving population will suffer deleterious mutations before they can reach the new peak. In effect, the population of the new protein is decimated by the sudden drop in available nonlethal mutations. If the population of proteins is able to cross this pathway then they have a new area of fitness landscape to explore, possibly improving fitness or acquiring new function. In this sense, a critical pathway is a bridge over a chasm, connecting two peaks on the evolutionary landscape.

ACKNOWLEDGMENTS

The authors are grateful for financial support from the Biotechnology and Biological Science Research Council in the form of a Bioinformatics Committee Studentship. The authors thank Professor Charles L. Brooks III of The Scripps Research Institute for his critical reading of the manuscript.

[1] A. R. Dinner, A. Sali, L. J. Smith, C. M. Dobson, and M. Karplus, Trends Biochem. Sci. **25**, 331 (2000).
[2] D. W. Miller and K. A. Dill, Protein Sci. **6**, 2166 (1997).
[3] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[4] S. Kumar, B. Ma, C.-J. Tsai, N. Sinha, and R. Nussinov, Protein Sci. **9**, 10 (2000).
[5] G. Tiana, R. A. Broglia, and E. I. Shakhnovich, Proteins: Struct., Funct., Genet. **39**, 244 (2000).
[6] D. M. Taverna and R. A. Goldstein, Biopolymers **53**, 1 (2000).
[7] E. Bornberg-Bauer and H. S. Chan, Proc. Natl. Acad. Sci. U.S.A. **96**, 10689 (1999).
[8] M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1983).
[9] J. D. Hirst, Protein Eng. **12**, 721 (1999).
[10] H. S. Chan and K. A. Dill, Proteins: Struct., Funct., Genet. **24**, 335 (1996).
[11] K. A. Dill, Biochemistry **29**, 7133 (1990).
[12] W. Kauzmann, Adv. Protein Chem. **3**, 1 (1959).
[13] S. Roy and M. H. Hecht, Biochemistry **39**, 4603 (2000).
[14] K. A. Dill, Protein Sci. **8**, 1166 (1999).
[15] E. I. Shakhnovich, Curr. Opin. Struct. Biol. **7**, 29 (1997).
[16] H. S. Chan and K. A. Dill, J. Chem. Phys. **95**, 3775 (1991).
[17] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, Protein Sci. **4**, 561 (1995).
[18] A. Irbäck and E. Sandelin, Biophys. J. **79**, 2252 (2000).
[19] M. D. Finucane, M. Tuna, J. H. Lees, and D. N. Woolfson, Biochemistry **38**, 11604 (1999).
[20] L. Li, L. A. Mirny, and E. I. Shakhnovich, Nat. Struct. Biol. **7**, 336 (2000).
[21] F. C. Berstein, T. F. Koetzle, G. J. B. Williams, F. Edgar, J. Meyer, M. D. Brice, O. Kennard, T. Shimanouchi, and M. Tasumi, J. Mol. Biol. **112**, 535 (1977).
[22] A. R. Ortiz and J. Skolnick, Biophys. J. **79**, 1787 (2000).
[23] A. Babajide, I. L. Hofacker, M. J. Sippl, and P. F. Stadler, Folding Des. **2**, 261 (1997).
[24] G. Giugliarelli, G. Micheletti, J. R. Banavar, and A. Maritan, J. Chem. Phys. **113**, 5072 (2000).
[25] K. Yue, M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **92**, 325 (1995).
[26] R. Backofen, S. Will, and E. Bornberg-Bauer, Bioinformatics **15**, 234 (1999).
[27] S. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford University Press, Oxford, 1993).