Functional Model Proteins: Structure, Function and Evolution

Benjamin P. Blackburne

A thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

February 2004

Contents

List of Figures

List of Tables

Refereed publications arising from this thesis

B.P. Blackburne and J.D. Hirst, *Evolution of Functional Model Proteins* J. Chem. Phys., *115*, 1935-1942 (2001).

N. Krasnogor, B.P. Blackburne, E.K. Burke, and J.D. Hirst. *Multimeme algorithms for protein structure prediction* In Parallel Problem Solving from Nature VII (PPSN-2002), volume *2439* of Lecture Notes in Computer Science, 769-778, Springer Verlag (2002).

B.P. Blackburne, and J.D. Hirst, *Three Dimensional Functional Model Proteins: Structure, Function and Evolution*, J. Chem. Phys. *119*, 6 3453-3460 (2003).

Abstract

An understanding of the forces of molecular evolution can yield insights into the nature of protein sequences and structures. In this thesis we use a minimalist model of proteins to provide a mapping from genotype to phenotype. The model is based on physically realistic forces of protein folding and includes a treatment of protein function. Thus we can find the fitness of a sequence from its ability to fold to a stable structure and perform a function. We can study the evolutionary landscapes occupied by the model proteins; the set of all sequences mapped on to their corresponding fitnesses and connected to their one mutant neighbours. Understanding the structure of these landscapes can help us understand both the origin of features of natural proteins and the scope for protein design.

We initially consider a two-dimensional minimalist model. Various features of the landscape are characterised. Longer chains are seen to possess wider functional diversity and form larger families of viable sequences connected by single point mutations. We observe that otherwise disconnected landscapes are linked together by single mutations, that we term critical edges.

We then consider a three-dimensional model. Prototype sequences are observed that can withstand many mutations and are highly stable. These sequences form a highly interconnected part of the landscape, and so offer scope for an increased rate of evolution. Real protein structures are often highly designable, with many sequences that fold to a single structure. The origins of this effect are considered.

Finally, we directly address the effect of landscape structure on evolution through population dynamics simulations. Populations are observed to move to a steady state population, the distrution of which can be predicted from the the structure of the landscape and the partition function. The implications of the distribution of populations on our landscapes are discussed.

Acknowledgments

Many thanks to my supervisor, Prof. Jonathan Hirst, for his direction, advice, criticism and depth of knowledge. I am greatful to Andrew Gilbert for advice and mathematical support, and to James Melville and Mark Oakley for proofreading. For useful discussions on lattice model algorithms I must thank Natalio Krasnogor. Richard Wheatley assessed my yearly reports, for which I am grateful.

A huge 'thank-you' to Beth for her help and support throughout the research period, and especially during the writing-up period. Thanks to Graham Mackenzie for letting me tap into his extensive Java knowledge. Ross Hunter, John McNeany, Darragh O'Neill, David Rogers, Tim Watson and Matt Wood have provided random helpful tidbits and evenings in the pub.

List of abbreviations

DNA Deoxyribonucleic Acid

RNA Ribonucleic Acid

TSE Transition State Ensemble

HP Hydrophibic/Polar (potential)

MJ Miyazawa-Jernigan (potential)

Chapter 1

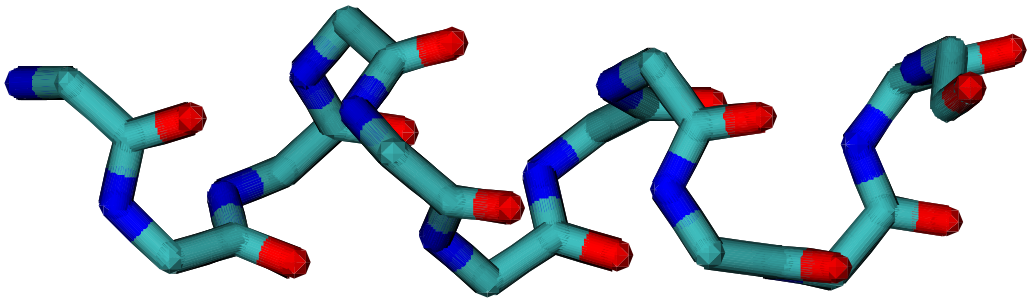
Protein Structure

1.1 How do Proteins Fold?

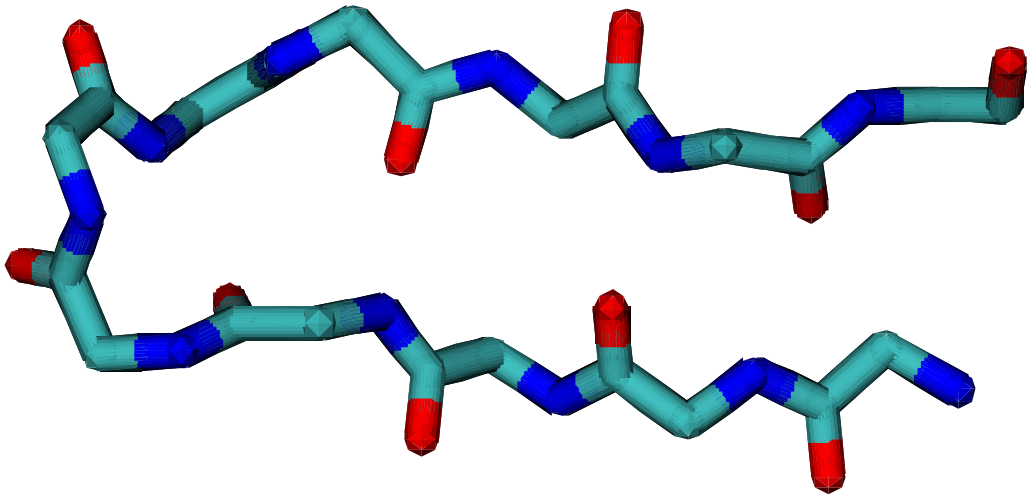
1.1.1 What is a protein?

Proteins are constructed from the twenty naturally-occurring amino acids. They are formed in the cell by the ribosomes, which build proteins from gene sequences that have been transcribed from DNA to RNA. A linear chain of amino acids (the "primary structure") is built that then folds into a specific ("native") three dimensional structure. The three dimensional structure is classified into the local order ("secondary structure"), such as the well known α-helices and β-sheet structures (see figure 1.1) and the "tertiary structure", the specific arrangement in which these structures come together. The folding of a protein sequence into its native state is thus a crucial mechanism in the expression of the cell's genomic instructions. While

the process of protein construction is well known, the process of protein folding is less well understood. However, much insight has already been achieved both through experiment and theory.



(a) α-helix



(b) β-sheet

Figure 1.1: Backbone structure of (a) an α-helix, and (b) a β-sheet, taken from the structure of lysozyme (PDB code 1HEL [36]).

What is the physical nature of the native state? The "thermodynamic hypothesis" of Anfinsen [2], that the native protein structure is the thermo-

dynamically stable structure, is now widely accepted. The alternative hypothesis, that structure is determined by processes in the cell, is certainly possible. Chaperonin proteins are present during the folding of many proteins, which may be interpreted to mean that a protein's native state is metastable, a local minimum enforced by conditions in the cell on folding.

However, chaperonins may be true catalysts, and so simply help a protein achieve its thermodynamically stable state more quickly. Anfinsen [2] randomised the covalent disulfide bridges in bovine pancreatic nuclease, and on return to native conditions the native structure and function were reacquired. The demonstration of thermodynamic reversibility for one protein does not prove that all protein native states are at the global free-energy minimum. However, in light of further supporting evidence [10] and theoretical studies [14], it seems reasonable to conclude that the thermodynamic hypothesis is likely to hold, at least for small globular proteins.

1.1.2 What are the forces that govern protein folding?

The different forces in protein folding are explored in a paper by Dill [10], that has since become a "citation classic", with 1389 published citations at this time. In this study the possible dominant forces for protein folding are examined as quantitatively as possible. The in-depth detail of the exact nature of these forces is beyond the scope of this study, which intends to deal with their implications for evolution rather than consider the forces themselves. A simplified discussion is given below.

First Dill considers electrostatic forces. The charge on a protein is sensitive

to changes in pH, due to the deprotonation of COOH groups at high pH, or protonation of NH_2 groups at low pH. At the iso-electric point the charges on a protein will be balanced; however, at extremes of pH a protein will contain a high net negative or positive charge. This net charge will have a non-specific repulsive effect resulting in the unfolding of the protein molecule, as the density of charge on an extended conformation is lower than in the compact native state. The destabilisation of proteins at extreme pH is well known and consistent with this effect [10].

A different effect could be caused by the interaction of specific charges on the protein chain. Could this be the "glue" that holds proteins together? When oppositely charged amino acids are close, this pairing will be energetically favourable. However, the dielectric constant in the interior of a protein is low, and so the burial of charges is energetically unfavourable. For this reason, most charges are seen on the surface of the protein.

However, despite extensive ion pairing on protein surfaces that can be shown to increase stability, a series of observations contradict the theory that it is a major force in protein folding. Firstly, the net transfer of a proton from a COOH group to an NH_2 group, causing subsequent formation of a salt bridge, should result in a loss in volume of 14ml mol per ion pair. Instead an increase in volume is observed on protein folding. Secondly, a reliance on ion pairing would leave the stability of a protein highly dependent on the pH and ionic strength of the solution. In fact, little dependence is seen near the iso-electric point. Finally, alteration of the charges in a protein has little effect, and examination of protein sequences shows that ion bridges are not highly conserved in evolution.

The structures of proteins contain extensive hydrogen bonding. For this reason we must consider the possibility that hydrogen bonding is the driving force for protein folding. The primary objection that can be raised is that this theory does not describe why the folded state is energetically favourable compared to the unfolded state. The free energy difference from the breaking of two protein-water hydrogen bonds and the formation of a protein-protein and a water-water hydrogen bond is not substantial. It must be highly enthalpically favourable, in order to overcome opposing entropic forces. The conclusion of this argument is *not* that hydrogen bonds are *not* important. A folded protein structure must contain many protein-protein hydrogen bonds, in order to compensate for the loss of hydrogen bonding with water. However, they cannot be responsible for the folding of a protein to the native state.

A well known factor in protein folding is the "intrinsic properties" of amino acids, the combination of forces that tend to favour particular local interactions and lead to conformational preferences of secondary structure, such as α helix, or β sheet. From the knowledge of the local sequence, predictions can be made of the secondary structure, with some success [19, 37]. If secondary structure can be accurately predicted by knowledge of local forces, then this implies that those forces are primarily responsible for the secondary structure. In fact, latest methods are between 70% and 80% accurate [19, 37]. However, even a success rate of 80% requires an information content of only 46% [10], as the classification into helix, sheet or coil has an a priori success rate of 1/3.

Experimental observations show that α-helices can be stable in solution,

with increased helix length leading to increased stability. However, the most frequent α-helices in proteins are short. Furthermore, it is difficult to account for how short-range forces can account for the long-range contacts that are needed to form β-sheets.

Cui *et al.* [9] use a hydrophobicity based model (an 18-mer HP lattice model, see section 1.2) to examine the frequency of short patterns within the viable model proteins found. The distribution of patterns is found to be significantly non-random, with common patterns responsible for helix-like and turn-like structures. It is interesting in this context to note that this stems from a model where local interactions (i.e. secondary structure propensities) are not explicitly included. Clearly local propensities are important, but they cannot be responsible for the collapse of a sequence to a specific native structure.

An additional force must be present that drives protein sequences from the molten globule [30] state of an ensemble of compact (secondary) structures to the native state. The possibility that hydrophobicity is this driving force for protein folding was argued by Kauzmann [21], whereby the formation of one "antihydrogen-bond", i.e. contact between two hydrophobic residues, would lead to the formation of a full hydrogen bond between two water molecules. Clearly the formation of extra hydrogen bonds is of much greater energetic importance than a small change in strength of hydrophobic interactions.

The mixing of oil and water gives a useful experimental model for this system. At 25°C the mixing of oil and water is opposed by entropy. The

mixing enthalpy is small and may even be favourable. This makes oil and water atypical - generally it would be expected that the entropy of mixing of two solutions would be positive. What is the molecular reason for this entropic effect in oil and water? One proposed model is that when the hydrophobic elements are dissolved in water, they restrict the number of orientations that the surrounding water molecules can adopt and still form hydrogen bonds. Rather than lose potential hydrogen-bonds, this leads to the formation of ordered cages of water, and so a lower entropy.

Further evidence for this mechanism comes from the large positive heat capacity of mixing, which is not observed for simple (non-hydrophobic) mixing. This means that the enthalpy and entropy of mixing are temperature dependent. At high temperatures, the opposing force is enthalpic, not entropic. This is consistent with the cage model proposed above. At high temperatures, the water molecules are able to occupy the previously inaccessible conformations, and the entropic effect is now favourable on mixing. However, this means that hydrogen bonds will be lost, leading to an unfavourable enthalpy upon mixing.

Understanding that hydrophobic forces are the main determinant of protein folding has led to advances in protein design. A “bury the grease” strategy of designing amphipathic secondary structural elements by binary-patterning of hydrophobic and hydrophilic residues can be used to design sequences that fold to target structures such as a four-helix bundle [28].

Simplified models of protein folding, based on only the hydrophobic effect as a folding code, have given great insight into mechanisms of protein folding.

Such models can show how the folding code may be primarily determined by the hydrophobic effect. The use of these models is the basis for this thesis, and they are discussed explicitly later (section 1.2).

1.1.3 The energy landscape

The concept of an energy landscape [12] stems from the so-called "new view" of protein folding. [4] Both the "old view" and the "new view" are attempts to explain protein folding in the face of the Levinthal Paradox [24]. In a protein chain backbone, there are two allowed bond rotations for each amino acid that can determine the shape of the protein. These are the ϕ and ψ dihedral angles. See figure 1.2. Levinthal reasoned [24] that for a protein chain of 150 amino acids, even if the ϕ and ψ angles were each restricted to only ten angles, there are 10^{300} possible conformations. Clearly, there is insufficient time for a folding protein to sample all these conformations, and so the search must be guided. The "old view" envisages a specific sequence of rearrangements occurring in each protein, as it follows a path to the folded structure. This folded structure may be the free-energy minimum, or it may be a metastable state.

The "new view" is an alternative explanation of how protein sequences can circumvent the exploration of a huge number of allowed conformations before folding. In this view, there is not a single path to the native state, but rather the random search is guided by an energetic bias. This bias is often conceptualised as a folding funnel (see figure 1.3). The folding funnel concept has been supported by and refined by simulations of protein models

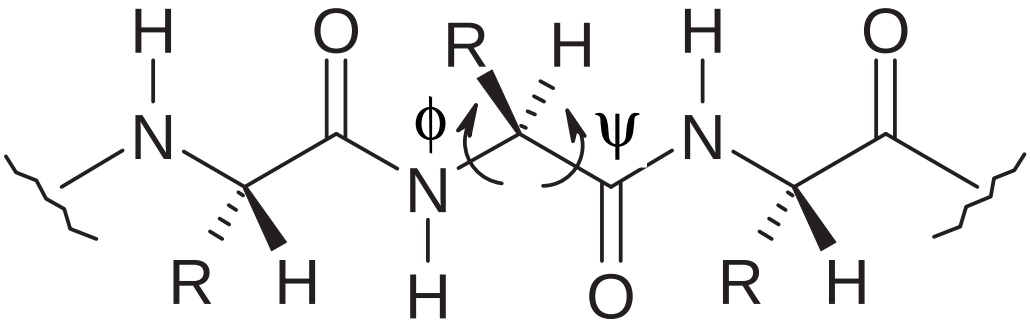


Figure 1.2: Part of a peptide backbone, showing the ϕ and ψ angles.

[7, 12] and experiment [12].

What is more important for determining the shape of a protein energy landscape: the sequence or the native structure? Somewhat surprisingly, it is possible that the rate of folding (which is strongly dependent on the shape of the landscape) can be almost entirely determined by the target structure [3]. An important factor determining the rate of protein folding is the type of long-range interactions required in the native state. It is more helpful, rather than thinking in terms of "local" and "non-local" interactions, to consider the "contact order" of a structure. This is the ratio of the average separation in the linear chain of those residues that are in contact in the native structure to the length of the chain. A high contact order structure will require that more distant residues are brought together in order for the native structure to be formed. This requires overcoming a greater amount of the opposing entropy that is the cause of the free energy barrier between the unfolded and folded states. The higher the contact order, the higher this barrier, and so the slower the rate of folding.

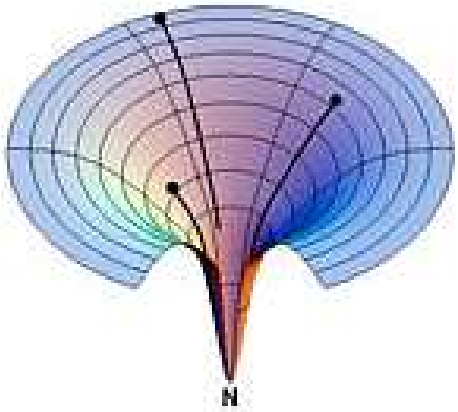


Figure 1.3: Illustration of an idealised folding funnel. Each point on the landscape represents a specific conformation, with similar conformations adjacent in the landscape. The height of the landscape represents the energy. The energetic bias present means that any structure on the landscape is able to fold to the native structure (the thermodynamic minimum). Figure taken (with permission) from reference [7]

That the sequence is less important than the structure seems counter-intuitive, but it is supported by some experimental evidence. Baker [3] cites three experimental approaches that vindicate this. One experimental study [22] concludes that the folding rate can be increased as often as decreased by random mutations. Variants of the IgG binding domain of peptostreptococcal protein L (62 residues, shown in figure 1.4) were made through the construction of phagemid libraries of mutants and mutagenesis of the non-IgG binding residues. Of 14 substantially changed sequences half increased the folding rate and half decreased it. The folding rates changed from $61s^{-1}$ for the wild type, to between 4 and $180s^{-1}$. We can conclude that the majority of mutations in proteins do not greatly affect the folding rate, and therefore are not altering the folding mechanism.

Furthermore, experiments to probe transition state ensembles of proteins that are sequentially unrelated but structurally similar reveal that the transition state is insensitive to the sequence. The folding rate of small proteins with two-state folding kinetics has been shown to be negatively correlated with the contact order [1], although this does not hold for larger proteins with three-state kinetics, where the rate correlates well with the chain length [13]. From this evidence, the topology of the native structure is shown to be important in determining the folding rate, but it is not the sole determinant [26].

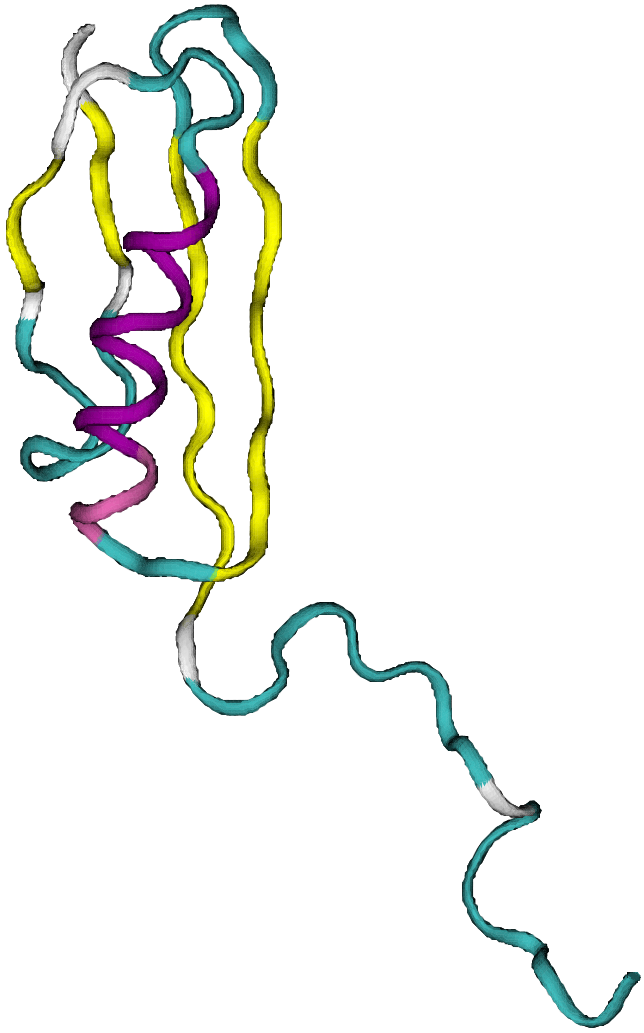


Figure 1.4: Ribbon representation of peptostreptococcal protein L. α-helices are drawn in purple, β-sheets in yellow. The figure was generated with the visualisation package VMD [17] from PDB file 2PTL [35].

1.1.4 Cooperative protein folding

The folding of small single-domain proteins is a two-state process [8]. Either sequences are in the fully unfolded state or the native state, with a tiny population of intermediate partially folded structures. Thus, protein folding is known as a cooperative, all-or-nothing transition. Although the situation is complicated for larger, multi-domain proteins, the individual domains can be shown to follow independent two-state processes [31].

The mechanism for this must be expressed by the shape of the free energy landscape. A landscape that contains a "peak" that must be climbed, before descending to the native state free energy minimum would lead to this two-state cooperativity [15]. As our understanding of the process of protein folding increases, we can begin to understand how the landscape has become structured to induce cooperative folding.

1.1.5 The nucleation mechanism of protein folding

The secondary structure of a protein refers to a level of local order that arises when a protein is folded. The protein chain is folded into the well known α-helices, β-sheets and turns. The tertiary structure is the three dimensional arrangement of amino acids. One way of considering the secondary and tertiary structures is that secondary structure results from "local" interactions (between amino acids near in the chain) and tertiary structure "long-range" interactions (those between distant regions of the chain). Here, long-range does not refer to the effective range within which the forces exert an effect,

but rather to the distance along the protein chain between interacting amino acids.

The process of protein folding leads to the formation of both of these types of interactions. However, an important question, which is relevant to the understanding of the forces in protein folding, is whether protein folding proceeds in a hierarchical fashion, with the formation of secondary structures, which then diffuse together to form the tertiary structure. This possibilty is the motivation for the diffusion-collision model of protein folding [20], which has recently been applied to the study of the folding of three-helix bundle proteins [18]. It assumes that a set of microdomains, which may be portions of secondary structure such as α-helicies, are in fast equilibrium between the native and denatured state. The folding process consists of the diffusion and coalescence of these microdomains into the native structure. It is also possible that the process involves collapse into the tertiary structure, which then drives formation of the elements of secondary structure. Current opinion states that hydrophobic collapse will naturally result in secondary structure that appears to be ordered, akin to the fitting together of nuts and bolts in a jar [10]. In this view, forces such as hydrophobic interactions and local energetic factors cause secondary structure formation from within an already highly restricted structure.

A recent review by Mirny and Shakhnovich [26] details the nucleation mechanism, a proposed model of protein folding. This defines a transition state ensemble (TSE) as a set of conformations that represent a free energy barrier to folding. The two-state cooperativity of protein folding is accounted for by this model. A protein must scale the free energy barrier of the TSE

by collapsing to a somewhat compact state. This is favourable in enthalpic terms, and is thought to involve the formation of a set of contacts that are present in the native state known as the folding nucleus [26]. However, clearly the collapse of a protein chain to the TSE involves a huge loss of entropy. As already stated it involves the formation of long-range (tertiary) contacts, and so the number of available conformations in the TSE is much lower than that of a fully unfolded chain. This model is somewhat agnostic about the formation of secondary structure - some secondary structural features may or may not be present in the TSE. However, in contrast to hierarchical models, it does involve long-range contacts.

This model offers a refinement of the folding funnel model. Although a simple energetic funnel will result from the successive accumulation of favourable native contacts, it is important to note that this funnel is enthalpic, and there are opposing entropic forces. Under the nucleation mechanism the initial folding process requires uphill movement on a free energy landscape as an unfavourable entropic process dominates a favourable enthalpy. Once past the TSE, the free energy gradient will be downhill, dominated by a favourable enthalpy.

It is interesting to consider this model in the context of the correlation of contact order with folding times, described earlier. This correlation may arise because the contact order of the folding nucleus and the native structure are well correlated - perhaps the contact order of the folding nucleus is a better determinant of folding times.

1.2 Lattice models of protein structure and folding

Full detail molecular simulations of proteins are feasible on modern computer hardware on nanosecond timescales or shorter. Typical proteins fold on millisecond or longer timescales. For this reason, over the past 30 years a series of reduced, so-called minimalist models of proteins have been developed to explore the effects of the major forces involved in protein folding. This section will look at the application of minimalist models to aspects of protein structure and folding. The following section will consider their application to questions in evolution.

Minimalist models, also known as lattice models, are designed to increase the detail in some aspects of proteins (complete sampling of all sequences or structures) at the expense of other aspects (fine molecular details). In fact, it may be advantageous to consider reduced models initially, simply because aspects of atomistic detail may begin to obscure the general principles that the simulations are designed to uncover [11]. Examples of lattice model proteins are shown in figure 1.5.

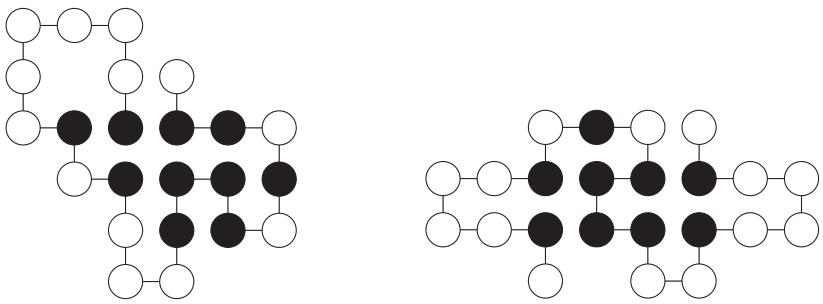


Figure 1.5: Two examples of lattice model proteins. Black beads are hydrophobes, white beads are polar.

To this end, Taketomi *et al.* developed a simplified model of protein folding [33] that has become known as the Gō model. In this model, individual amino acids are represented as beads on a chain, which is restricted to lie on a square lattice. This means that the chain is two-dimensional and bond angles between beads are only allowed to be multiples of 90°. The chain is self-avoiding - no conformation is allowed where two beads occupy the same lattice point. The energy of a conformation is determined from the set of contacts, beads that are nearest neighbours on the lattice in that particular conformation, but not nearest neighbours in the chain.

In the Gō model the native structure is specified in advance. From this, energy interactions are specified to be favourable (those that occur in the native state) or neutral. A second model is introduced, where the contacts are specified as before, but additional random favourable interactions are added, so that 1/3 of all interactions are favourable. The comparison of these two models with a model where all interactions are favourable allows the effect of the specificity of the interactions to be assessed.

The specification of favourable contacts as those present in the native state separates this model from the majority of models studied later. In order to incorporate greater realism later models include a physically realistic potential; however the Gō model has certain advantages. The conclusions of the study are not dependent on a specific model of the interactions of protein folding, but rather are general for polymers that fold to a specific state. The model displays highly cooperative folding, that can be considered more realistic than some recent models, even when favourable non-native interactions are included. Finally, the advance knowledge of the native state

is advantageous if the protein's energetic landscape is to be explored without an exhaustive search through all possible structures.

The folding of the models was simulated using the Metropolis Monte Carlo method [25]. A spike in heat capacity on heating, two-peaked population profiles, and sigmoidal denaturation curves were observed for the protein-like models, indicating an all-or-nothing transition. This established that protein-like features can emerge from a highly simplified model. That the fine molecular details of protein sequences are not necessary for the general features of protein folding is critical in establishing the minimalist model as a useful tool for biophysics.

A shift from Gō models, which focus on the interactions from a known native state, to the HP model and its descendants came in a study by Lau and Dill [23]. The insight that the driving force for protein folding is hydrophobicity, as discussed earlier, leads to the formulation of the folding code in those terms. Proteins are modelled as the simplest possible heteropolymer, with only two types of residues. Each bead on the chain is assigned to be either hydrophobic, H, or polar, P. This increases scope of the study from the Gō models; instead of asking if protein-like features can arise from a simplified lattice model this asks if those features can arise from a lattice model with a hydrophobic folding code.

Lau and Dill computed the folding behaviour of short (eleven residues or less) HP model proteins. The "standard" HP model sets interactions between two hydrophobic nearest neighbours to be favourable (negative in energy), while other (H-P, P-P) interactions are energetically neutral. In

this study, the partition function is used to determine the density of states for a set of sequences of fixed length when the energy of an H-H contact ($-\varepsilon$) is varied. For higher values of ε, corresponding to stronger interactions, enthalpy dominates and a compact ("native") conformation is often observed, depending on the sequence composition. For lower ε, the effect of entropy means that sequences are unfolded. Only a very few sequences are shown to have a single native state. Of those sequences with degenerate native states, the structures of those states are, on average, quite different. Although longer sequences could not be examined exhaustively, they were considered through sampling of compact conformations, and a higher proportion were thought to have unique native states.

A later study by Chan and Dill used lattice models to examine the shape of protein energy landscapes [6]. A longer chain with 13 residues was used along with the "standard" HP model, where H-H contacts are favourable and other contacts neutral. The folding of 173 13-mer HP sequences with non-degenerate ground states was studied with two different move sets. The landscape was characterised as rugged; the lower in energy a conformation is, the more likely it is necessary to go uphill on the landscape to reach another conformation of equivalent energy. A perfectly smooth funnel (figure 1.3) would not possess this feature. This ruggedness does not depend on consideration of entropy; it is produced directly from the enthalpy of the structures.

A Metropolis [25] method of folding was applied to the sequences, and two stages of folding were identified. The initial process was rapid collapse to an ensemble of compact conformations with buried hydrophobes. The second

process involved a slow search for the native state, that involved a movement over energetic barriers. A class of fast folding sequences was identified, where this slow reconfiguration was faster, due to the presence of more paths that avoid kinetic traps.

A wide variation in time taken to fold was observed, depending on the move-set used. The implications of this are that it is difficult to draw conclusions about the nature of the kinetic processes in protein folding directly from Monte Carlo simulations, and also that algorithms that attempt to find protein native states from Monte Carlo methods should have their move-sets chosen carefully.

Another important result from this study is the presence of mutations that increase the speed of folding by three orders of magnitude by lowering the energy barrier between compact non-native conformations and the native state. A simple characteristic of the landscape of rapid folding stable sequences is noted. Sequences that are fast folders and have stable native states are those with a large "energy gap", the difference in energy between the native state and the lowest energy non-native conformations.

A common modern lattice model implementation is a 27-mer on a cubic lattice. The advantage of this model is that the compact conformations are restricted to a three by three by three cube, and its 103,346 conformations unrelated by symmetry can be easily enumerated. The disadvantage of this model is that the cubic lattice is six-coordinate rather than four-coordinate, leading to a much greater number of possible extended conformations for a given length. Exhaustive enumeration of all conformations of the 27-mer

on a cubic lattice is, to our knowledge, not attempted; instead Monte Carlo methods are used.

The concept of an energy gap is dealt with in a study by Sali *et al.* [32]. The 27-mer model on a cubic lattice was employed, and the requirements for folding were tested. The interaction parameters were drawn from a Gaussian distribution, rather than the HP model. A set of 200 random sequences were tested, and 30 were shown to be good folders, reaching the ground state in four of ten Monte Carlo trials. A large energy gap between the ground and first excited states, as well as between the ground and the denatured states, was shown to differentiate between the good and bad folders.

An extended cubic lattice model was employed to examine the folding nucleus theory [26]. Side chains were added to the model in an attempt to increase its realism. A side chain is represented as a bead attached to each bead in the backbone chain and occupying an adjacent lattice point. The interaction energies between the side-chains were those specified by Miyazawa and Jernigan [27]. The MJ potential is derived from a database of protein crystal structures. The frequency of contact pairs between amino acids is used to calculate a contact energy. These energies are ideal for use in lattice simulations, where a representation of the twenty-letter amino acid alphabet is desired.

A single sequence with a single native state was examined. The folding nucleus contacts were identified as those contacts usually formed during folding when 41% of native contacts were formed. This set of 11 native and five non-native contacts was compared with a control set of 11 random

native and five random non-native contacts. A structure containing the folding nucleus was able to fold in less than 1% of the number of steps of a random coil, whereas a structure containing the control set was on the same order of magnitude as a random coil. A mutant where the non-native nucleus contacts were repulsive resulted in a sequence that folded significantly more slowly, although the stability was not significantly affected.

1.3 Protein folding algorithms using lattice models

The simplified nature of lattice models make them a useful first step in the development of algorithms which attempt to predict the native structure of a protein from its sequence. If an algorithm is unable to find the optimum structure for a short chain in discretised space, then its applicability to real protein sequences must be in doubt. To this end, genetic algorithm [34] and Monte Carlo [5] based procedures have been developed to fold long lattice model chains. Other groups have developed more complex methods for constructing low energy structures for sequences [38, 39]. An alternative approach is to include the lattice model as a component of a folding algorithm [16] in order to reduce the conformational space that needs to be explored. The four-coordinate diamond lattice was chosen for this study by Hinds and Levitt as a reasonable compromise between conformational flexibility (it is three-dimensional) and complexity (it is four-coordinate).

Finally, lattice models can be used as a first step in protein design studies. A paper by Pokarowski *et al.* examines a sequence designed to fold into an Greek-key anti-parallel β-sheet structure [29] on the 12-coordinate face

centred cubic lattice. Two types of interactions were found to be necessary and sufficient to cause a full cooperative transition to this native state, avoiding possible misfolds. Firstly, the long range hydrophobic interaction of the HP model, and secondly, a local secondary structure propensity for β-sheet or random coil. This is interesting from the point of view of protein design, where we need to consider the minimum requirements for folding to a structure in order to obviate the complexity of the features of the twenty-letter amino acid alphabet. Some success has been achieved with the elegant two-letter alphabet approach [28], designing sequences that fold to a four-helix bundle. It is more difficult to design sequences that fold well to a β-sheet structure using this two-letter code [28], and perhaps a computational lattice approach could lead to additional insight.

In the following chapter, we consider the characteristics of molecular evolution, and the application of lattice models to their study.

Bibliography

[1] E. Alm and D. Baker. Matching theory and experiment in protein folding. *Curr. Opin, Str. Biol.*, 9:189–196, 1999.

[2] C.B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[3] D. Baker. A surprising simplicity to protein folding. *Nature*, 405:39–42, 2000.

[4] R. L. Baldwin. Protein folding. Matching speed and stability. *Nature*, 369:183–184, 1994.

[5] U. Bastolla, H. Frauenkron, E. Gerstner, P. Grassberger, and W. Nadler. Testing a new Monte Carlo algorithm for protein folding. *Proteins*, 32:52–66, 1998.

[6] H.S. Chan and K. Dill. Transition states and folding dynamics of proteins and copolymers. *J. Chem. Phys.*, 100:9238–9257, 1994.

[7] H.S. Chan and K.A. Dill. Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics. *Proteins*, 30:2–33, 1998.

[8] T.E. Creighton. The problem of how and why proteins adopt folded conformations. *J. Phys. Chem.*, 89:2452–2459, 1985.

[9] Y. Cui, W.H. Wong, E. Bornberg-Bauer, and H.S. Chan. Recombinatoric exploration of novel folded structures: A heteropolymer-based model of protein evolutionary landscapes. *Proc. Natl. Acad. Sci. U. S. A.*, 99:809–814, 2002.

[10] K.A. Dill. Dominant forces in protein folding. *Biochemistry*, 29:7133–7155, 1990.

[11] K.A. Dill, S. Bromberg, K.Z. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, and H.S. Chan. Principles of protein-folding - a perspective from simple exact models. *Protein Sci.*, 4:561–602, 1995.

[12] A.R. Dinner, A. Sali, L.J. Smith, C.M. Dobson, and M. Karplus. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.*, 25:331–339, 2000.

[13] O.V. Galzitskaya, S.O. Garbuzynskiy, D.N. Ivankov, and A.V. Finkelstein. Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. *Proteins*, 51:162–166, 2003.

[14] S. Govindarajan and R.A. Goldstein. On the thermodynamic hypothesis of protein folding. *Proc. Natl. Acad. Sci. U. S. A.*, 95:5545–5549, 1998.

[15] M.H. Hao and H.A. Scheraga. Molecular mechanisms for cooperative folding of proteins. *J. Mol. Biol.*, 277:973–983, 1998.

[16] D.A. Hinds and M. Levitt. A lattice model for protein-structure prediction at low resolution. *Proc. Natl. Acad. Sci. U. S. A.*, 89:2536–2540, 1992.

[17] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *J Mol Graph*, 14:33–38, 1996.

[18] S.A. Islam, M. Karplus, and D.L. Weaver. Application of the diffusion-collision model to the folding of three-helix bundle proteins. *J. Mol. Biol.*, 318:199–215, 2002.

[19] D.T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292:195–202, 1999.

[20] M. Karplus and D. L. Weaver. Diffusion-collision model for protein folding. *Biopolymers*, 18:1421–1437, 1979.

[21] W. Kauzmann. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.*, 14:1–63, 1959.

[22] D.E. Kim, H. Gu, and D. Baker. The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc. Natl. Acad. Sci. U. S. A.*, 95:4982–4986, 1998.

[23] K.F. Lau and K.A. Dill. A lattice statistical-mechanics model of the conformational and sequence-spaces of proteins. *Macromolecules*, 22:3986–3997, 1989.

[24] C. Levinthal. How to fold graciously. *Mossbauer Spectroscopy of Biological Systems*, pages 22–24., 1969.

[25] N. Metropolis and Ulam S. The monte carlo method. *J Am Stat Ass*, 44:335–341, 1949.

[26] L. Mirny and E. Shakhnovich. Protein folding theory: From lattice to all-atom models. *Annu. Rev. Biophys. Biomolec. Struct.*, 30:361–396, 2001.

[27] S. Miyazawa and R.L. Jernigan. Estimation of effective interresidue contact energies from protein crystal-structures - quasi-chemical approximation. *Macromolecules*, 18:534–552, 1985.

[28] D.A. Moffet and M.H. Hecht. De novo proteins from combinatorial libraries. *Chem. Rev.*, 101:3191–3203, 2001.

[29] P. Pokarowski, A. Kolinski, and J. Skolnick. A minimal physically realistic protein-like lattice model: Designing an energy landscape that ensures all-or-none folding to a unique native state. *Biophys. J.*, 84:1518–1526, 2003.

[30] O.B. Ptitsyn and V.N. Uversky. The molten globule is a 3rd thermodynamical state of protein molecules. *FEBS Lett.*, 341:15–18, 1994.

[31] G. Ramsay and E. Freire. Linked thermal and solute pertubation analysis of cooperative domain interactions in proteins. Structural stability of diptheria toxin. *Biochemistry*, 29:8677–8683, 1990.

[32] A. Sali, E. Shakhnovich, and M. Karplus. Kinetics of protein-folding - a lattice model study of the requirements for folding to the native-state. *J. Mol. Biol.*, 235:1614–1636, 1994.

[33] H. Taketomi, Y. Ueda, and N. Gō. Studies on protein folding, unfolding and fluctuations by computer simulation. i. the effect of specific amino acid sequence represeneted by specific inter-unit interactions. *Int. J. Peptide Protein Res.*, 7:445–459, 1975.

[34] R. Unger and J. Moult. Genetic algorithms for protein folding simulations. *J. Mol. Biol.*, 231:75–81, 1993.

[35] M. Wikstrom, T. Drakenberg, S. Forsen, U. Sjobring, and L. Bjorck. 3-dimensional solution structure of an immunoglobulin light chain-binding domain of protein-L - comparison with the IgG- binding domains of protein-G. *Biochemistry*, 33:14011–14017, 1994.

[36] K. P. Wilson, B. A. Malcolm, and B. W. Matthews. Structural and thermodynamic analysis of compensating mutations within the core of chicken egg white lysozyme. *J Biol Chem*, 267:10842–10849, 1992.

[37] M.J. Wood and J.D. Hirst. Predicting protein secondary structure by cascade-correlation neural networks. *Bioinformatics (in press)*, 2004.

[38] K. Yue and K.A. Dill. Forces of tertiary structural organization in globular-proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 92:146–150, 1995.

[39] K. Yue, K.M. Fiebig, P.D. Thomas, H.S. Chan, E.I. Shakhnovich, and K.A. Dill. A test of lattice protein-folding algorithms. *Proc. Natl. Acad. Sci. U. S. A.*, 92:325–329, 1995.

Chapter 2

Models of Protein Evolution

2.1 Adaptive and Neutral Evolution

Modern protein sequences are a consequence of the process of mutation and selection. The constraints imposed on protein sequences by selection are the ability to fold on a biologically relevant timescale, to possess a stable native structure, and to perform a biological function. In a theoretical treatment of protein evolution, the fitness of a sequence is a term that covers all of these features.

After a mutation the resulting protein sequence may be less fit (often), more fit (rare), or the difference in fitness may be insignificant. Mutations that do not affect the fitness of a sequence are termed neutral mutations, and the importance of these mutations was first identified by Kimura [10], and independently, King and Jukes [11, 8]. The fixation of these mutations is termed "neutral drift". Because of the process of neutral drift, one should

be careful not to consider that all features of proteins are explained by adaptive evolution. Understanding how the two processes of adaptive and neutral evolution have contributed to the features of modern proteins is an area of modern inquiry, and is the subject of this thesis.

2.2 Landscapes

We have already considered energy landscapes as a description of protein folding. Landscapes are also a useful concept for the study of evolution. Fitness landscapes, (or evolutionary landscapes, the terms are used interchangeably in this thesis) were introduced by Sewall Wright [22], who considered evolution to be a process of movement on a landscape towards areas of higher fitness. This hill climbing metaphor is inverted from the concept of moving to lower energy on a protein folding landscape.

For the purposes of molecular evolution we can consider that the variables of the landscape constitute sequence space, the set of all possible sequences for a protein of a given length. The "height" of each sequence is its fitness. A single step on the landscape corresponds to a single point mutation. In this way, proteins can be considered to evolve by successive fixation of point mutations that carry the population to a sequence of higher fitness. This process neglects effects such as recombination, but point mutations are an important component of evolution, and as such, this concept is a useful first step.

An important aspect of any theoretical treatment of molecular evolution is

the determination of the structure of the fitness landscape. This requires a method for determining a phenotype for each genotype. The *NK* model [9] used a random assignment of fitnesses to produce a tunably rugged landscape. While this model has given great insight into the effect of landscape structure on processes such as protein evolution, it does not address the question of how the physical constraints of protein structure and function affect the landscape.

As discussed in the previous chapter, although the process of protein folding is complex, much insight can be gained from the use of minimalist models. In the same way, lattice models can be used to produce a mapping from genotype to phenotype. The landscapes produced by these methods are themselves simplified, but still depend on realistic physical characteristics of proteins, allowing hypotheses to be generated based on the current understanding of protein structure and function.

2.3 Designability

One important characteristic of landscapes that is ideally addressed by lattice models is "designability". Designability stems from an observation made of real known protein structures, that there are certain folds or classes of folds that account for a large number of sequences [4], sometimes with less than 30% sequence identity [19]. Other structures are under-represented [19]. The designability of a structure is the number of sequences that fold to that structure as their native state, i.e. the number of sequences that "design" that structure. Accounting for the variation in designability is a

current area of inquiry, and minimalist protein models are ideally suited to examining this question, due to the necessity to consider a large number of sequences.

The study of designability is tied up with other questions. Important for the consideration of theories of pre-biotic protein formation is the question of what proportion of sequences fold to a compact globular structure. The distribution of designability amongst structures is also important for theories of convergent and divergent evolution; highly designable structures may be encoded by such a large fraction of sequence space that convergent evolution is rendered highly likely.

An early study by Lau and Dill [12] addresses the probability of randomly choosing a sequence that encodes for a single chosen structure. This is a rephrasing of the question of designability. From estimates of the number of sequences with a unique native state they conclude that the majority of compact structures will be encoded for by a large number of different sequences. A later study by Chan and Dill [3] examines this by exhaustive enumeration of HP sequences of length 14. They demonstrate that the more compact the structure, the higher the designability (note that designability is referred to as convergence in this study). However, this study neglects direct consideration of whether a sequence folds to one or more lowest energy structures, i.e. whether a sequence has a single native state. They do observe, however, that sequences that fold with low degeneracy fold to more compact structures, and that more compact structures are more highly designable than less compact structures.

A more direct examination of sequences with non-degenerate native structures was made by Lipman and Wilbur [15]. Exhaustive enumeration of sequences up to length 19 were made, with a folding sequence taken to be one that has a single native structure. Enumeration was made of only compact and nearly compact structures. The designability (referred to as the size of an "equivalence class" in the paper) shows that while the majority of structures that are native states are only encoded for by a few (e.g. $<$ 10) sequences there are a minority that are encoded for by more sequences (e.g. $>$ 50). This study [15] demonstrated how levels of designability are distributed amongst structure space in a Zipf-like fashion [23]. Zipf's Law is an observation, originally made of word frequency in languages, that the frequency of occurrence of some event (P) can be found from a function of its rank (r), $P_r = \frac{1}{r^a}$, where a is close to unity. In a Zipf-like distribution some structures are highly represented in sequence space, just as some words (such as "the" and "of" in English) are highly represented in English.

A study by Li *et al.* [13] compared the characteristics of highly designable structures with less designable compact structures. A small fraction (0.12%) of the compact structures was shown to occupy a separate class of extremely highly designable structures. This class of structures was also shown to be, when averaged over their designing sequences, more thermodynamically stable than those of less designability. These structures were seen to possess secondary structure like features, such as bundles of pleats and long strands comparable to α-helices and β-sheets in real proteins. Some of these structures were also shown to have tertiary symmetries.

One possible explanation for the correlation between designability and aver-

age thermodynamic stability is that, for a given sequence folding into a given structure, the more stable the sequence is in that structure, the more mutations can be made without causing a competing near-native structure to become equal or lower in energy, and so remove the ability of that sequence to fold.

What is the origin of the symmetry often seen in highly designable structures? One possible explanation is that highly designable structures have a large number of surface to core transitions - regions of the chain when bead i is on the surface, and bead $i + 1$ is in the core. These transitions lead to a geometrical regularity which may be the source of symmetry. A recent investigation by Wang *et al.* [20] uses an elegant HP solvation model to investigate the contribution made by surface to core transitions to symmetry. A statistical analysis of this model revealed a correlation that was insufficient to explain the symmetries shown in highly designable structures. Another possible explanation for the symmetry is that the highly designable structures are constructed from a small set of highly designable substructures. The argument is that highly designable substructures are "winning solutions", the duplication of which would create symmetry.

A different approach was taken by Govindarajan and Goldstein [6], who determined the optimal set of interactions for each member of the set of maximally compact 27-mer structures on the cubic lattice. They employed a measure of structural stability called foldability. The structures which can be optimised to the highest foldabilities are also those that were most designable. This is consistent with the above explanations of designability; the structures that are the most stable allow the most mutations, and so

are the most designable.

The study of designability has applications outside of protein evolution. For example, if real protein structures are frequently highly designable, it may be helpful to limit structure prediction algorithms to the highly designable structures, possibly by construction from known, highly designable motifs. Furthermore, the more designable natural protein structures are, the simpler the energy potential that can be used in attempting to fold a particular sequence. If structures are highly designable, this means that a great many mutations can be made to a sequence before it loses its ability to fold. Therefore specific individual interactions between amino acid pairs are not as important as in a protein structure that is designed by only a few sequences.

The effect of choice of alphabet on the designability of protein structures has been investigated by Buchler and Goldstein [2], and Li *et al.* [14]. The first study contrasted the structures found using the 20-letter Miyazawa-Jernigan (MJ) potential [16] against that found for several variants of the HP model. It concludes that the structures found to be highly designable under the HP model (those with high levels of symmetry) were not the same as those found using the MJ potential. The second study concluded that the designable structures found were highly sensitive to the MJ potential chosen. The original Miyazawa and Jernigan paper [16] describes several possible potentials depending on the conditions chosen. Li *et al.* found that the use of the e_{ij} potential from reference [16] results a similar set of structures being found for both the HP and MJ potentials. The e_{ij} MJ potential gives a relative energy made up of the formation of a contact between residues i and j, combined with the removal of the i-water and

j-water contacts. This energy is dominated by the hydrophobic interaction, and is claimed to be the most suitable for this type of study [14]. Other potentials described by Miyazawa and Jernigan exclude this interaction.

The representation of designable structures under evolving populations was directly studied on a square lattice model with the MJ energy potential [16] constrained to a 25-mer square by Taverna and Goldstein [17]. The designability of the structures of this model is compared with their occupancy under a population dynamics simulation. The occupancy of structures during the simulation cannot be easily predicted - many populations are higher or lower than might be expected from their designabilities. This means that the structure of the surrounding landscape has a great influence over the population dynamics. The sequences encoding for a structure may have highly fit neighbours, and will receive population from mutations to those neighbours . Other sequences may not have fit neighbours, and so their populations will be relatively lower. On average the distribution of structures was skewed, with highly designable structures more likely to be occupied than would be expected from a simple count of their designability.

2.4 Evolutionary landscapes

The specific structure of an evolutionary landscape is crucial in determining its effect on populations. Lattice models are capable of addressing this question, by moving away from the generalised questions of designability, and considering the fine structure of the landscape they produce. In one of the first studies to use lattice models to examine protein evolution, Lau

and Dill [12] examine the effect of mutations on a set of 251 13-mer HP sequences that are good folders. The study examines the effect of single and double point mutations. For the sequences in the study, the majority of mutations (single and double) are predicted to be neutral. The majority of non-neutral mutations take place in the core rather than on the surface of the protein. In the case of an adaptive mutation, the effect is cooperative, i.e. the change in structure usually involves the reconfiguration of many bonds.

Lipman and Wilbur [15] more directly describe some of the basic features of the landscapes occupied by lattice model proteins. A fit sequence is defined as one that folds to a non-degenerate native state, and adaptive mutations are those that change the structure of the native state. This study revealed characteristics of the landscape that hold true for longer chains and full enumeration of structure space.

Lipman and Wilbur observe disconnected landscapes that encode for identical structures. This observation is consistent with theories of convergent evolution. An important observation is that the ratio of neutral to adaptive mutations is greater than one. This is one indicator that sequence space is structured such that there are areas where, in order for adaptive evolution to take place, neutral drift must first take place. For this reason, we must be careful when analysing the evolutionary record not to misinterpret the data. There may exist neutral mutations that are wholly necessary before adaptive evolution can take place, but these mutations must not be misinterpreted to be adaptive themselves. One implication of this is that the need for neutral drift in between periods of adaptive evolution can act as a

"brake" on evolution, as the so-called hill-climbing process must stop for a period of stochastic neutral drift.

2.4.1 Structure of evolutionary landscapes

Increased computational power has allowed more recent studies to characterise evolutionary landscapes in greater detail. Exhaustive enumeration in sequence and structure space means that landscapes can be characterised through the sizes and structures of "neutral nets", sequences that are connected by single point mutations in sequence space and which all fold to a single structure.

Bornberg-Bauer and Chan consider the standard HP model with the density of states calculated by exhaustive enumeration of all possible structures [1]. The effect of the standard "protein-like" HP model is compared with the AB model. In the AB model, like attracts like, i.e. A-A and B-B interactions are favourable. The AB model is useful as a control offering an insight into the origin of the features of the landscape. Features that result from the HP model but not from the AB model may be ascribed to the more protein-like nature of the hydrophobic model.

Bornberg-Bauer and Chan observe an interesting structure to the neutral nets formed by the HP model. The most thermodynamically stable sequence is also the sequence with the most allowed non-lethal point mutations. This sequence is known as the prototype sequence. The structure of the surrounding neutral net shows a funnel-shape; each layer of mutations away from the prototype sequence is less stable than the previous layer. Put an-

other way, the prototype sequence has the highest stability, and so occupies the bottom of a funnel. It is surrounded by a circle of neighbours that are less stable, and so higher up the funnel. The next layer (those two mutations from the prototype sequence) are less stable than the first layer. This funnel structure appears to be a more frequent characteristic of the HP model than the AB model.

2.4.2 Evolution of model proteins

A tool for analysis of the characteristics of landscapes is the simulation of the process of evolution on their surfaces made through population dynamics. A population of sequences is evolved on a landscape according to a set of rules. These rules can be deterministic or stochastic. Stochastic models have the advantage that multiple runs can offer further insight, due to the random element, and the inclusion of randomness may more accurately represent evolution. However, the need to make many runs in order to draw solid conclusions makes them more computationally intensive. Furthermore, deterministic models, by their design, simulate large (effectively infinite) populations that may be more relevant to large populations of microorganisms. Here, in keeping with the nature of model proteins, population dynamics simulations are not used to make quantitative predictions but rather to gain an insight into the general rules that govern populations on landscapes.

A study by Taverna and Goldstein [18] addressed the observation that proteins are only marginally stable. This aspect of protein thermodynamics was discussed in chapter 1. Most globular proteins possess a $\Delta G_{folding}$ of

only around -10 kcal/mol. This marginal stability has been suggested to be an adaptation - the increased flexibility that it confers may be necessary for optimal function. But as already discussed, care must be taken before ascribing to adaptive evolution something that can be explained by the neutral theory [10]. Taverna and Goldstein recall Dr. Pangloss from Voltaire's *Candide*, who considers that the nose is an obvious adaptation for the wearing of spectacles. Can the marginal stability of proteins be explained without adaptation?

Taverna and Goldstein perform population dynamics simulations on a two dimensional lattice model protein. Three populations of sequences were evolved together stochastically with the requirement that each population maintains a stability within a low, medium or high range of $\Delta G_{folding}$. These populations are allowed to evolve together, but the descendants of each group still selected for each group's original requirements. Eventually two populations die out, leaving one population as the "fixed" population. Of 25 trials, 24 resulted in the population with the low stability requirements being fixed. In the other trial, the population with medium stability became fixed.

These results can be easily understood if evolutionary landscapes are structured as superfunnels [1]. Three populations were set to evolve against each other. One population was constrained to the centre of the superfunnel, one population to the outside. Another was constrained to an in-between area. The population which is on the most evolutionary stable (least thermodynamically stable) landscape, that is, the landscape with the most sequences and the most allowed mutations between those sequences, will be the most

likely to survive. Thus, the outer (least thermodynamically stable) population will increase relative to the other populations, as they cannot grow as quickly due to more lethal mutations. Although individual sequences at the centre of the superfunnel will be expected to be more highly populated than individual sequences at the rim [1], these results suggest that there will be a higher total population at rim due to the many more sequences there [18].

There is another way of picturing this process. On average, there are more neutral mutations leading towards the rim of the superfunnel than to the centre. On average, moving from the centre to the rim, each layer of the superfunnel will contain more sequences than the last. A population undergoing neutral evolution will be likely to move towards the rim of the funnel, where stability is marginal. In section 5.7.3, we consider the effect of this evolutionary landscape topology on evolution for a function.

Evolutionary landscapes conceptually place great importance on point mutations. Population dynamics simulations are an obvious technique for the extension of evolutionary landscape to the study of recombination. Depending on the restrictions applied to the definition of a stable folded protein, sequence space may be fragmented into several landscapes of viable sequences disconnected by one or more lethal mutations. Recombination could provide a mechanism for crossing over to an otherwise disconnected landscape.

A paper by Cui *et al.* [5] addresses this issue though the study of the population dynamics of HP lattice model proteins. For a chain of length 18, they find 6,349 viable sequences in 700 networks, with the largest network containing 4,553 sequences. About 28% of crossover events are shown to lead

to at least one viable sequence. The fitness of a sequence is defined from the stability of its native state. A deterministic model of evolution is employed, albeit with the stochastic element of a fluctuating environment. Variation of this stochastic factor determines the selective pressure, and more rapidly changing environments are shown to encourage a broader exploration of sequence space. The exploration of sequence space is also broadened by an increased incidence of recombination. However, recombination works best at amplification of the diversity of a larger set of sequences generated by point mutations, and is poor at generating diversity from a small set of sequences. Thus, a sizable point mutation rate is still favourable to generate a set of sequences for use in recombination.

2.4.3 Including biological function in lattice models

The previous models are designed to give a mapping from genotype to phenotype using a realistic physical basis. The fitness (phenotype) measures given usually depend on the stability of the sequence, or a two-state fit/not-fit measure. A more realistic, biological interpretation of fitness depends on the ability of a protein to perform a given function. Clearly all foldable proteins are not equally fit in a given role, and the fitness of a protein does not depend, or at the very least depends only partly on, the stability of the native state. Therefore it may be useful to examine biological function as an aspect of protein evolution independent of the requirements for folding and stability.

Williams *et al.* devised a lattice-based model of protein structure and func-

tion that addresses this issue [21]. They consider 16-mer lattice proteins with the native state of a maximally compact conformation on a four by four square lattice, using the twenty letter MJ potential [16]. The fitness of a sequence is determined from the ability of a tetra-peptide ligand to bind to one of the sides. A stochastic population dynamics simulation is used to compare the evolution of 1000 sequences with selection for compactness of the native state and for fitness (ligand binding).

Under both criteria compact structures quickly appear which are retained for the rest of the simulation. This rapid freezing-in is due to the effective increased selective pressure that results from the population of sequences becoming fitter. The effect of this rapid freezing-in is to reverse the previous observation that occupancies were highly skewed [17] (see section 2.3). Instead, structures are selected quickly, and a structure of low designability now represents a small island of fitness that, at a low mutation rate, is difficult for the population to leave.

The following chapters detail our model, which includes an explicit definition of function, and the conclusions we can draw about the nature of evolutionary landscapes. Chapter 3 deals with a two-dimensional model, extended from earlier work [7]. We call the fit sequences from this model "functional model proteins" due to their separate consideration of function. Chapter 4 examines a set of three-dimensional functional model proteins. Chapter 5 uses population dynamics simulations to examine the evolution of functional model proteins explicitly.

Bibliography

[1] E. Bornberg-Bauer and H.S. Chan. Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. *Proc. Natl. Acad. Sci. U. S. A.*, 96:10689–10694, 1999.

[2] N.E.G. Buchler and R.A. Goldstein. Universal correlation between energy gap and foldability for the random energy model and lattice proteins. *J. Chem. Phys.*, 111:6599–6609, 1999.

[3] H.S. Chan and K.A. Dill. Sequence space soup of proteins and copolymers. *J. Chem. Phys.*, 95:3775–3787, 1991.

[4] C. Chothia. Proteins - 1000 families for the molecular biologist. *Nature*, 357:543–544, 1992.

[5] Y. Cui, W.H. Wong, E. Bornberg-Bauer, and H.S. Chan. Recombinatoric exploration of novel folded structures: A heteropolymer-based model of protein evolutionary landscapes. *Proc. Natl. Acad. Sci. U. S. A.*, 99:809–814, 2002.

[6] S. Govindarajan and R.A. Goldstein. Why are some protein structures so common? *Proc. Natl. Acad. Sci. U. S. A.*, 93:3341–3345, 1996.

[7] J.D. Hirst. The evolutionary landscape of functional model proteins. *Protein Eng.*, 12:721–726, 1999.

[8] T. H. Jukes. The Neutral Theory of Molecular Evolution. *Genetics*, 154:956–958, 2000.

[9] S. Kauffmann. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, Oxford, 1993.

[10] M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, 1983.

[11] J. L. King and T. H. Jukes. Non-Darwinian evolution. *Science*, 164:788–798, 1969.

[12] K.F. Lau and K.A. Dill. A lattice statistical-mechanics model of the conformational and sequence-spaces of proteins. *Macromolecules*, 22:3986–3997, 1989.

[13] H. Li, R. Helling, C. Tang, and N. Wingreen. Emergence of preferred structures in a simple model of protein folding. *Science*, 273:666–669, 1996.

[14] H. Li, C. Tang, and N.S. Wingreen. Designability of protein structures: A lattice-model study using the Miyazawa-Jernigan matrix. *Proteins*, 49:403–412, 2002.

[15] D.J. Lipman and W.J. Wilbur. Modeling neutral and selective evolution of protein folding. *Proc. R. Soc. Lond. Ser. B-Biol. Sci.*, 245:7–11, 1991.

[16] S. Miyazawa and R.L. Jernigan. Estimation of effective interresidue contact energies from protein crystal-structures - quasi-chemical approximation. *Macromolecules*, 18:534–552, 1985.

[17] D.M. Taverna and R.A. Goldstein. The distribution of structures in evolving protein populations. *Biopolymers*, 53:1–8, 2000.

[18] D.M. Taverna and R.A. Goldstein. Why are proteins so robust to site mutations? *J. Mol. Biol.*, 315:479–484, 2002.

[19] S.A. Teichmann, C. Chothia, and M. Gerstein. Advances in structural genomics. *Curr. Opin. Struct. Biol.*, 9:390–399, 1999.

[20] T.R. Wang, J. Miller, N.S. Wingreen, C. Tang, and K.A. Dill. Symmetry and designability for lattice protein models. *J. Chem. Phys.*, 113:8329–8336, 2000.

[21] P.D. Williams, D.D. Pollock, and R.A. Goldstein. Evolution of functionality in lattice proteins. *J. Mol. Graph.*, 19:150–156, 2001.

[22] S. Wright. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the Sixth International Congress on Genetics*, pages 354–366, 1932.

[23] G.K. Zipf. *Psycho-Biology of Languages*. Houghton-Mifflin, 1935.

Chapter 3

Two Dimensional Functional Model Proteins

3.1 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 [10], 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 [21]. Various types of binding behaviours have been explored using a modification of the two-dimensional HP lattice model. The standard model of hydrophobic H and polar P residues [19] was extended to include a third, ligand L, monomer type [21]. 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 [18].

Evolutionary aspects of proteins are increasingly investigated using lattice models [26, 25, 4]. 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 [17] 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 [26] 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 con-

fined to a two-dimensional square lattice [25]. The fitness of sequences was taken to be related to the ability of the protein to fold. In another characterisation of evolutionary landscapes [4], 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 has been briefly discussed previously [13], 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 [13], in which Hirst 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 be viable. These include the requirement for a unique, non-degenerate lowest energy ground state. To impose cooperative folding, we require an energy gap between the native state and the first excited state. Thus, functional 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. Hirst has previously studied chains of up to length 20 [13]. Here, we present a more detailed characterisation 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. 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 [6], 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 [13, 6], this leads to an interaction energy matrix of the form:

$$E = \begin{pmatrix} -2 & 1 \\ 1 & 1 \end{pmatrix} \tag{3.1}$$

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

In minimalist models, the protein chain is restricted to lie on a lattice. This discretised model permits the exhaustive enumeration of all possible conformations. We employ the two-dimensional square lattice. As noted by Dill [8], 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 characterise 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 [13]. A contact map, a list of nearest-neighbour nonbonded monomers, is constructed from each walk. It is not necessary to evaluate the energy for every conformation [6], 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.

3.2 Methods

3.2.1 Minimalist models

Part of this chapter (sections 3.2.1 though to 3.3.2) details work undertaken for an undergraduate degree at the University of Nottingham. This work is included here as the results provide a necessary context for the further studies and the methods described were applied to later studies.

Various amino acid alphabets may be used in minimalist models of proteins [6]. One of the simplest, the HP model [19], derives from the insight that the hydrophobic interaction is a major determinant of protein folding [7, 16]. This view has found widespread theoretical and experimental support, and continues to be useful in guiding experimental studies [23]. In the HP model, the interaction energy between two hydrophobic H residues is ε, 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.

3.2.2 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 two 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, they are a reasonable basis for a minimalist model of proteins. A functional model protein is required to have a non-degenerate 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 behaviour of real proteins. We chose one of several possible simple criteria to model cooperativity, which if not universally accepted, does have some support [24]. We discuss this further in section 4.2 (page 93). 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 [5, 9], 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 analysed. 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 se-

quence 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 favourable 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 analysed for 19-mers, 20-mers, and 21-mers. Data for the latter are presented in Table 3.1, which shows the frequency of occurrence of functional model proteins with respect to the number of contacts and sequence composition. The shading in Table 3.1 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.

3.2.3 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 pro-

Table 3.1: 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

tein 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 [13]. In this case, the most hydrophobic pocket is used to compute the fitness of the functional model protein. Pockets may be characterised 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 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 characterised. 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 analysed. 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 analyse 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.

3.3 Results

3.3.1 Sequence and structural characterisation

Our results and discussion focus primarily on chains of length 19 to 23. Functional model proteins of lengths 11 to 20 have been studied previously [13]. 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 3.2, 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 non-degenerate 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 to 21 were analysed. The 21-mers have the broadest range in composition, containing between five 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 five H residues or more than 15. Thus, restricting the enumeration

Table 3.2: Conformational space of chains on the square lattice, and number of viable functional model proteins found

Number of monomers in chain	Number of conformations	Number of contact maps	Number of singly-degenerate contact maps	Number of functional model proteins
19	15 582 342	363 010	140 708	1 936
20	41 889 573	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

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 3.2), 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 Figure 3.3.1. It is evident that functional model 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 [14].

3.3.2 Function and evolutionary characterisation

The data in Table 3.3 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 favour 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 diver-

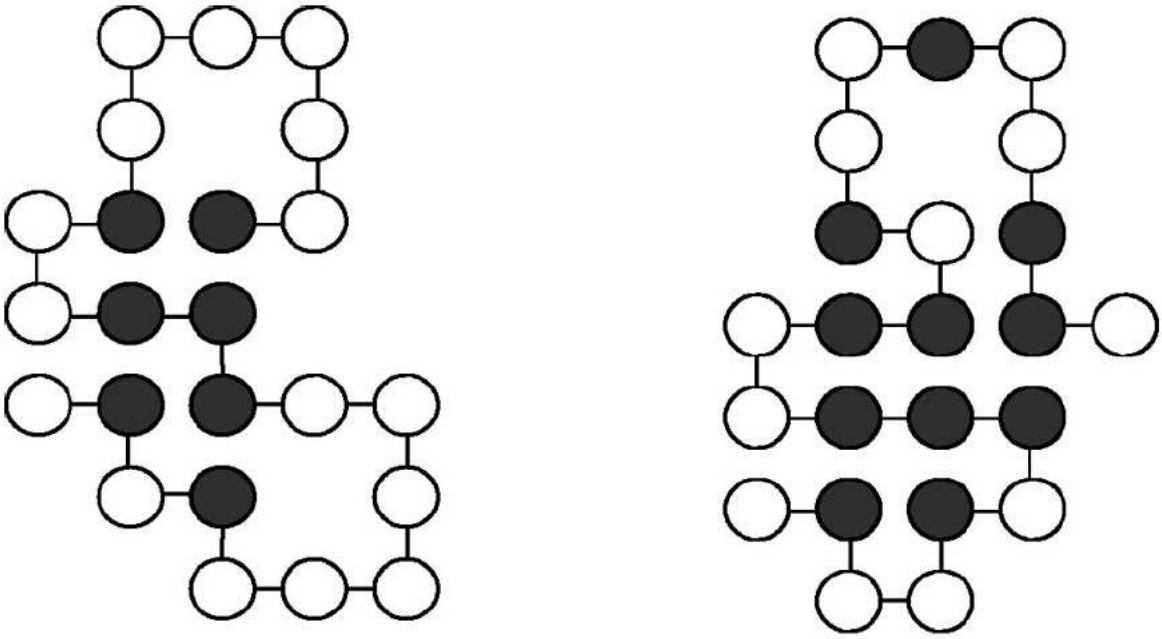


Figure 3.1: Two representative examples of functional model proteins on the square lattice. Black monomers are hydrophobic; white monomers are polar.

sity. 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.

Table 3.3: Normalised distribution of sequence composition of binding pockets. Fraction of sequences with f residues in the binding pocket.

	Chain length				
f	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

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. The sizes of the largest families and the numbers of functional model proteins found are given in Table 3.4. Several other properties of the evolutionary landscapes are given in Table 3.5. The table includes some new data on the shorter chains, and some data that have been corrected from an earlier study [13]. The sizes of the largest families increase with chain length. Most families have symmetry-related analogues

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 previous work [13] beyond 20-mers. The 23-mer produces many large families. There are 26 families of 23-mers with more than 100 members each.

Table 3.4: Numbers of functional model proteins

	No. functional model proteins	
n	Total	Largest Family
16	289	18
17	652	30
18	819	14
19	1 936	59
20	3 953	576
21	7 113	71
22	10 605	80
23	31 146	713

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 char-

Table 3.5: Characteristics of evolutionary landscapes.

	Mean allowed mutations/sequence		Neutral:Adaptive mutations		Mean number of binding pockets	
n	All seqs	Largest Family	All seqs	Largest Family	All seqs	Largest Family
16	1.42	1.28	1.34	0.53	1.05	1.67
17	1.90	1.70	2.15	0.82	1.06	1.13
18	1.28	2.29	5.92	0.78	1.05	1.14
19	1.86	3.29	2.61	1.49	1.13	1.14
20	1.74	3.19	2.77	1.72	1.12	1.16
21	1.87	3.43	2.81	1.77	1.14	1.14
22	1.41	3.05	3.43	1.71	1.17	1.33
23	2.39	3.98	2.73	2.51	1.21	1.21

acterised 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 characterise evolutionary landscapes more fully, one needs to consider not only the numbers of nonlethal mutations, but also their nature.

3.3.3 Structural and functional adaption

Functional model proteins, with their explicit definition of function, allow the characterisation 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 3.6 demonstrates that in practice there is a clear distinction in our model between mutations which affect structure and those which affect function.

Table 3.6 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

Table 3.6: Characterisation of adaption in terms of structure or function expressed as the fraction of nonlethal single point substitutions.

Chain Length	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

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 [17].

3.3.4 Critical edges

Graphs of the largest families of 21-mers, 22-mers, and 23-mers are depicted in Figures 3.2, 3.3, and 3.4. 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 Figures 3.2 and 3.4. Several consecutive critical edges form a critical pathway, which sometimes connects two different neutral networks, thus controlling evolution between

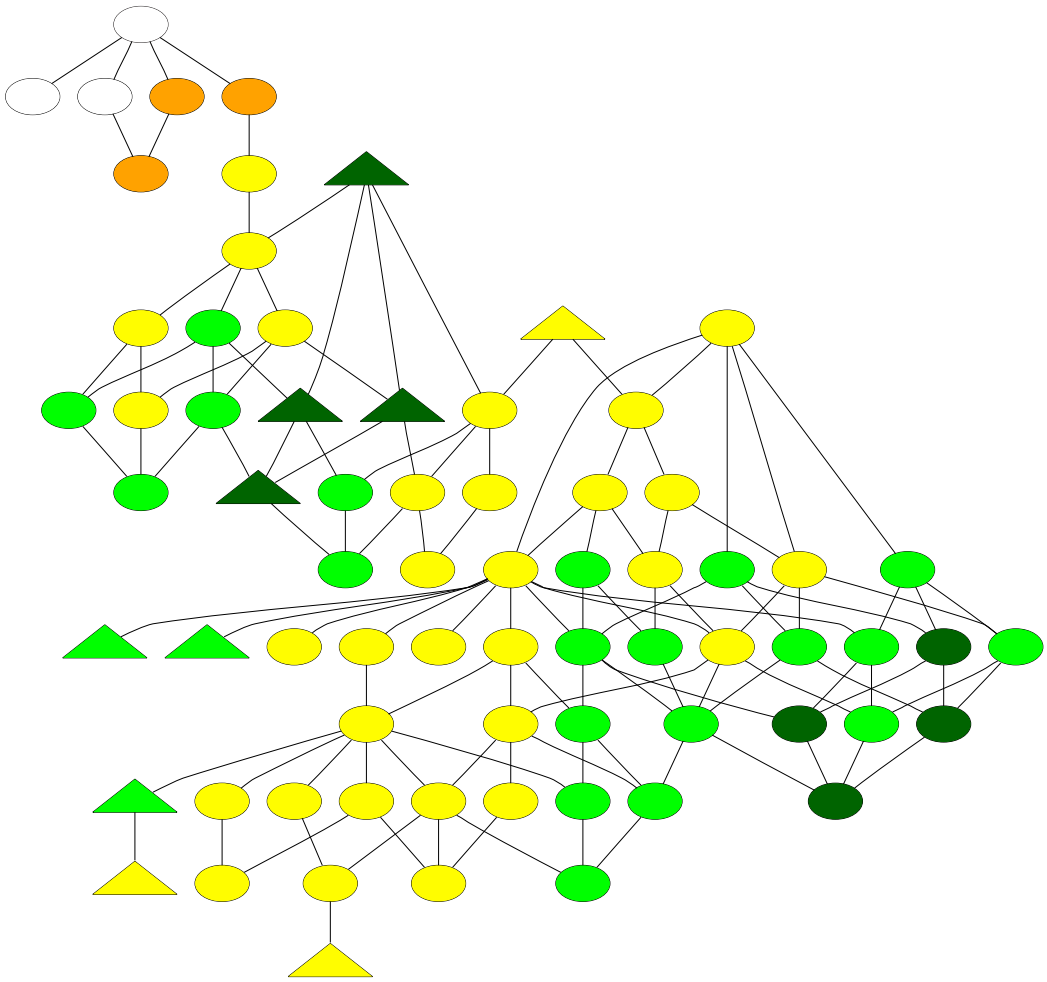


Figure 3.2: 71-member family of 21-mers. Nodes correspond to functional model proteins; edges correspond to single point mutations connecting two viable sequences. Nodes are colour coded based on function (the number of H residues in the binding pocket): white-1, orange-2, yellow-3, light green-4, dark 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 text.

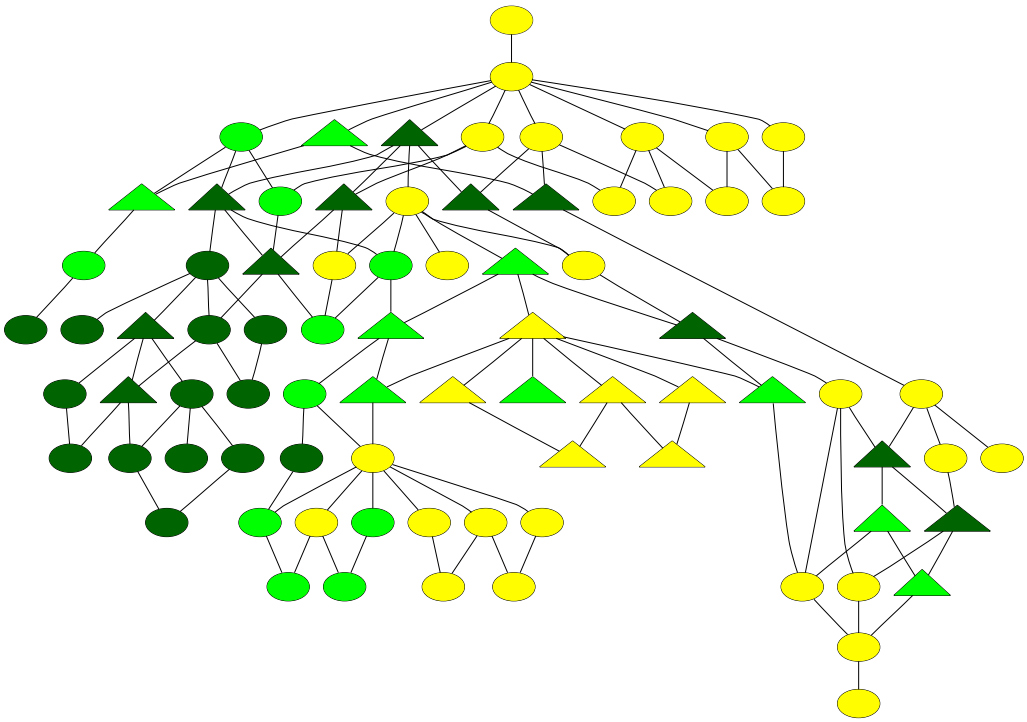


Figure 3.3: 80-member family of 22-mers, drawn with the same convention as figure 3.2.

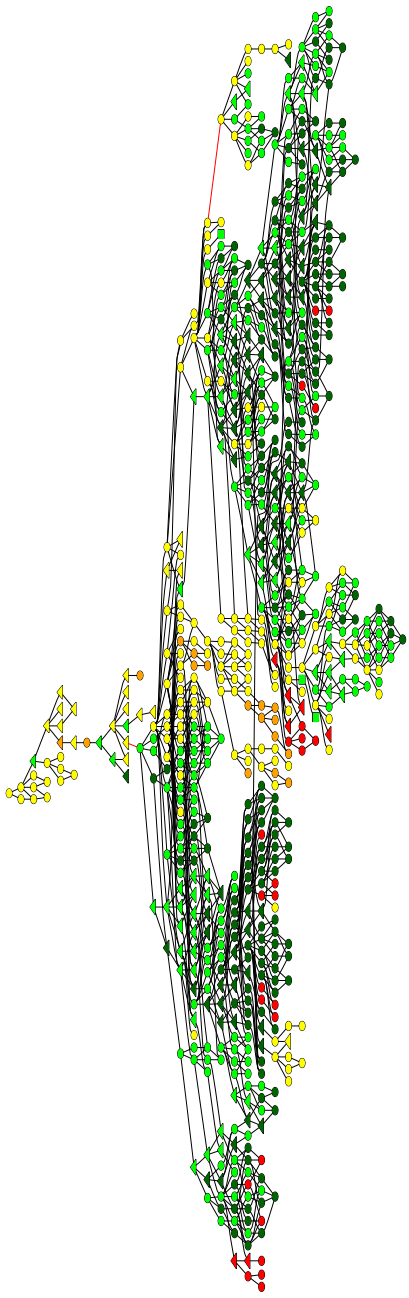


Figure 3.4: 713-member family of 22-mers, drawn with the same convention as figure 3.2.

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 seven 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 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 visualised 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.

3.3.5 Insertions and deletions

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 (Figure 3.5). 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.

3.3.6 Characterisation of surface pockets

We include an analysis of just the surface binding sites for chains of length 19 to 23. This is a subset of the sequences considered so far, and a corresponding decrease in family sizes is observed in Table 3.7. 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 (Table 3.8) closely follows the pattern already seen in Table 3.8. 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 (Figure 3.6). This is also shown in the elevated ratio of neutral:adaptive mutations. This could be because surface binding pockets are by definition small and at the edge of the protein and so 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

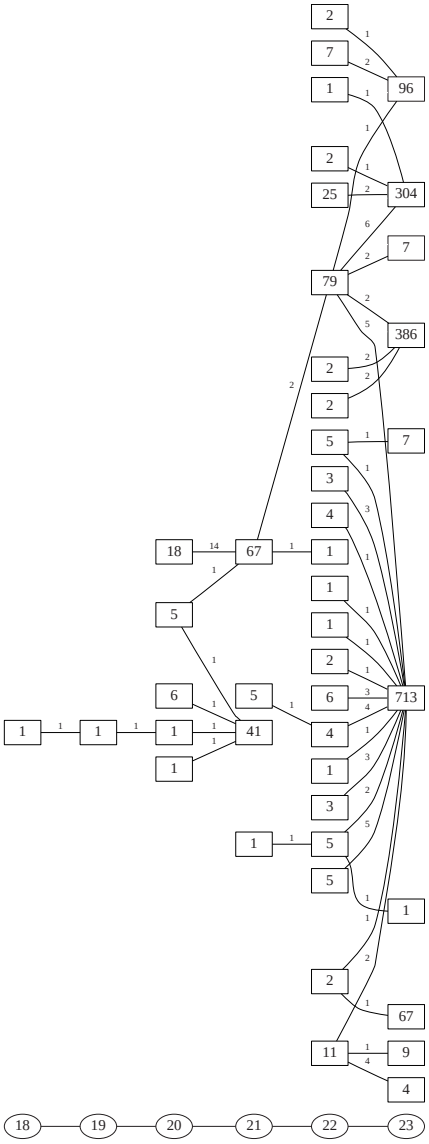


Figure 3.5: 1915-member family. The elliptical nodes show the chain length. Each rectangular node is a subfamily, labelled with the number of members. Each edge indicates one or more possible insertions/deletions, and is labelled with the number of connections between each family.

often consist of remote sections of the chain brought together. In proteins with surface pockets 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.

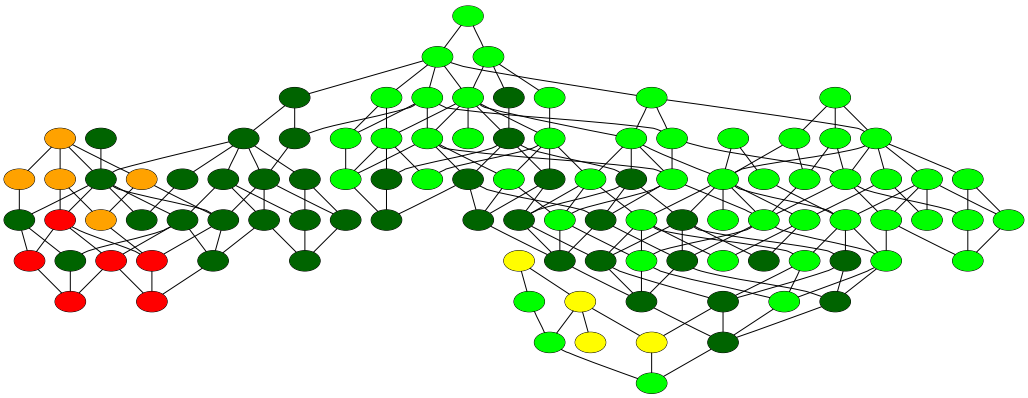


Figure 3.6: 105-member family of 23-mers with surface pockets, drawn with the same convention as Figure 3.2.

Table 3.7: Numbers of surface pocket functional model proteins

	No. functional model proteins	
n	Total	Largest Family
19	240	39
20	445	17
21	1 842	54
22	2 592	56
23	7 464	105

Table 3.8: Characteristics of evolutionary landscapes of surface pocket proteins.

	Mean allowed mutations/sequence		Neutral:Adaptive mutations		Mean number of binding pockets	
n	All seqs	Largest Family	All seqs	Largest Family	All seqs	Largest Family
19	1.98	2.92	2.07	1.59	1.01	1
20	1.34	2.47	3.52	9.5	1.01	1
21	1.91	2.89	3.66	4.57	1.05	1
22	1.47	3.5	5.24	4.44	1.06	1
23	1.81	3.94	3.87	4.45	1.08	1

3.4 Conclusions

In this chapter, we have characterised 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 [11].

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 nonnative 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 [26], which is also observed in an analysis of structures taken from the Protein Data Bank [3]. 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. [1] used a statistical measure of native-like 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 non-polar proteins [23], half demonstrated cooperative thermal denaturation, some of which were also fully monomeric in solution, leading to the conclusion that they display native-like folding.

Bornberg-Bauer and Chan [4] 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 favourable, exhibits a more fragmented landscape. The AB model provides a useful reference point, although it is widely re-

garded 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 [12] 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 [27, 2]. 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 modelled 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 [15], 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.

Bibliography

[1] A. Babajide, I.L. Hofacker, M.J. Sippl, and P.F. Stadler. Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. *Fold. Des.*, 2:261–269, 1997.

[2] R. Backofen, S. Will, and E. Bornberg-Bauer. Application of constraint programming techniques for structure prediction of lattice proteins with extended alphabets. *Bioinformatics*, 15:234–242, 1999.

[3] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data bank. *Nucleic Acids Res.*, 28:235–242, 2000.

[4] E. Bornberg-Bauer and H.S. Chan. Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. *Proc. Natl. Acad. Sci. U. S. A.*, 96:10689–10694, 1999.

[5] H.S. Chan and K.A. Dill. Sequence space soup of proteins and copolymers. *J. Chem. Phys.*, 95:3775–3787, 1991.

[6] H.S. Chan and K.A. Dill. Comparing folding codes for proteins and polymers. *Proteins*, 24:335–344, 1996.

[7] K.A. Dill. Dominant forces in protein folding. *Biochemistry*, 29:7133–7155, 1990.

[8] K.A. Dill. Polymer principles and protein folding. *Protein Sci.*, 8:1166–1180, 1999.

[9] K.A. Dill, S. Bromberg, K.Z. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, and H.S. Chan. Principles of protein-folding - a perspective from simple exact models. *Protein Sci.*, 4:561–602, 1995.

[10] A.R. Dinner, A. Sali, L.J. Smith, C.M. Dobson, and M. Karplus. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.*, 25:331–339, 2000.

[11] M.D. Finucane, M. Tuna, J.H. Lees, and D.N. Woolfson. Core-directed protein design. I. An experimental method for selecting stable proteins from combinatorial libraries. *Biochemistry*, 38:11604–11612, 1999.

[12] G. Giugliarelli, C. Micheletti, J.R. Banavar, and A. Maritan. Compactness, aggregation, and prionlike behavior of protein: A lattice model study. *J. Chem. Phys.*, 113:5072–5077, 2000.

[13] J.D. Hirst. The evolutionary landscape of functional model proteins. *Protein Eng.*, 12:721–726, 1999.

[14] A. Irbäck and E. Sandelin. On hydrophobicity correlations in protein chains. *Biophys. J.*, 79:2252–2258, 2000.

[15] S. Kauffmann. *The Origins of Order: Self-Organization and Selection in Evolution.* Oxford University Press, Oxford, 1993.

[16] W. Kauzmann. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.*, 14:1–63, 1959.

[17] M. Kimura. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, 1983.

[18] S. Kumar, B.Y. Ma, C.J. Tsai, N. Sinha, and R. Nussinov. Folding and binding cascades: Dynamic landscapes and population shifts. *Protein Sci.*, 9:10–19, 2000.

[19] K.F. Lau and K.A. Dill. A lattice statistical-mechanics model of the conformational and sequence-spaces of proteins. *Macromolecules*, 22:3986–3997, 1989.

[20] L. Li, L.A. Mirny, and E.I. Shakhnovich. Kinetics, thermodynamics and evolution of non-native interactions in a protein folding nucleus. *Nat. Struct. Biol.*, 7:336–342, 2000.

[21] D.W. Miller and K.A. Dill. Ligand binding to proteins: The binding landscape model. *Protein Sci.*, 6:2166–2179, 1997.

[22] A.R. Ortiz and J. Skolnick. Sequence evolution and the mechanism of protein folding. *Biophys. J.*, 79:1787–1799, 2000.

[23] S. Roy and M.H. Hecht. Cooperative thermal denaturation of proteins designed by binary patterning of polar and nonpolar amino acids. *Biochemistry*, 39:4603–4607, 2000.

[24] E.I. Shakhnovich. Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.*, 7:29–40, 1997.

[25] D.M. Taverna and R.A. Goldstein. The distribution of structures in evolving protein populations. *Biopolymers*, 53:1–8, 2000.

[26] G. Tiana, R.A. Broglia, and E.I. Shakhnovich. Hiking in the energy landscape in sequence space: A bumpy road to good folders. *Proteins*, 39:244–251, 2000.

[27] K. Yue, K.M. Fiebig, P.D. Thomas, H.S. Chan, E.I. Shakhnovich, and K.A. Dill. A test of lattice protein-folding algorithms. *Proc. Natl. Acad. Sci. U. S. A.*, 92:325–329, 1995.

Chapter 4

Three Dimensional Functional Model Proteins

4.1 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 [19] 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 visualising the process of molecular evolution is Wright's fitness landscape [34]. 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 modelling 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. Kauffmann [16] 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 [35], 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 protein-like 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 the stability of the protein. We define fitness as the number of hydrophobes in the binding pocket of the protein. This is an easily characterised physical property that is relevant to a possible function, the non-specific 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, whilst maintaining physical relevance. The use of lattice models in protein folding is well established [27, 10], and their application to

evolutionary problems is increasing. In one study, the origins of designability have been examined using a 20-letter lattice model [29]. In another study [33], 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 [28]. The natural propensity for evolution to select sequences for robustness to mutation due to the effects of quasispecies [11] was demonstrated. Each sequence is able to mutate into a neighbour in sequence space, and so the population at each point 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 [5]. 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 [9, 17]. Each residue is reduced to a single point on the lattice, and is assigned to be either hydrophobic or polar [5]. The power of this concept is illustrated by *de novo* protein design experiments, which select sequences with certain patterns of hydrophobic and polar residues [24]. 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 [23]. 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 [7]. Elements of secondary structures also show distinctive HP patterns. In one study [22], 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 [2] with the HP model. Sequences in a neutral net are shown to be organised such that they are centred around a prototype sequence, with stabilities increasing towards the centre. This feature is observed in our model, which is three-dimensional and uses a different folding code. 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 [20].

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 utilise a three-dimensional (3D) lattice as a step to more realistic structures. Unlike many other studies [12, 25, 26] which employ a maximally compact 27-mer on a cubic lattice, restricted to a 3x3x3 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 [26] has suggested that two letter models of proteins give rise to a set of designable structures that is different to that produced by larger alphabets, and that conclusions about designability drawn from studies of HP models have limited generality. However, as noted elsewhere [15] this suggestion itself rests on studies of maximally compact 5x5 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 [36]. It was suggested that the HP model may not display protein-like 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 below, on the diamond lattice using the shifted-HP model [5] show a large number of non-degenerate ground states with energy gaps, which also possess a binding pocket.

4.2 Methods

The HP model assumes a favourable interaction energy when two hydrophobes are in contact on the lattice but are not neighbours in the chain ($E_{HH} = -1$), whilst other contacts are neutral ($E_{HP} = E_{PP} = 0$). 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 [5]. In this case, $E_{HH} = -2$, and $E_{HP} = E_{PP} = 1$.

Functional model proteins have been previously studied on the square lattice [14, 1] (chapter 3). In this chapter we extend our analysis to the 3D diamond lattice. The diamond lattice is expected to incorporate more realism into our model, whilst maintaining the four co-ordinate 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 4.1, we show how conformational space grows with chain length. The data agree with early enumerations by Wall and Hioe [30] for chains up to length 21. The earlier enumeration ignores symmetry and rotation, so the number of conformations reported earlier are $24\Omega - 12$, where Ω is the number reported as "number of conformations" in Table 4.1. The growth in number of conformations is similar to the growth seen on the square lattice [1]; 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 Figure 4.1.

Table 4.1: Conformational space of chains on the diamond lattice.

Number of monomers in chain	Number of conformations	Number of contact maps	Number of non-degenerate 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

We define a viable sequence as one which satisfies several conditions. It must have a non-degenerate 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

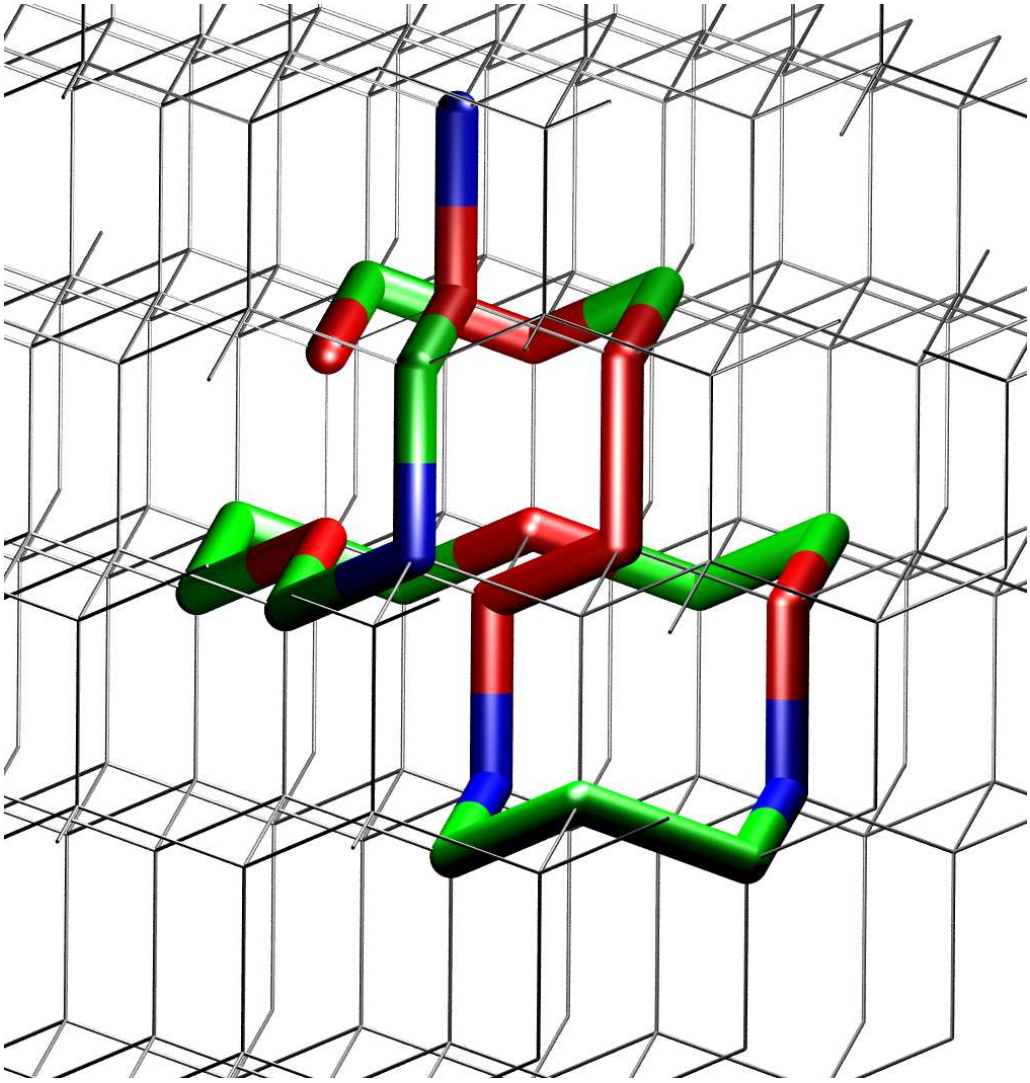


Figure 4.1: An example of a 25-mer diamond lattice protein. Red residues are hydrophobic, blue are polar. Green residues are subject to variation within the family. There are 151 sequences that fold to this structure.

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 [4]. 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 non-additive terms in the energy formulation [18]. Cooperativity may be an evolved feature [21] dependent on side-chain packing and other features [13, 6]. With respect to the overall aim of studying models with protein-like phenomenology, the energy gap requirement is useful for excluding model sequences with unprotein-like behaviour. More detailed work on cooperativity in lattice models would be desirable, but is beyond the scope of this thesis.

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 characterisation 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 [14]. To evaluate a HP sequence we consider the distribution of its energies over all

possible conformations. As previously noted [5, 14, 1], 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 [1] (Figure 3.5). 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 sequences 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 non-specific hydrophobic

binding as a model of fitness. Previously we defined the fitness, f, as the number of hydrophobic residues around the binding pocket [14, 1] (see page 53). On the square lattice, including next nearest neighbours 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 neighbours and next nearest neighbours, (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.

4.3 Results & Discussion

4.3.1 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, whilst 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 4.2 shows that the proportion of neutral mutations increases with increasing length and landscape size. Larger landscapes will stabilise 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 non-lethal 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.

Table 4.2: 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.

	All sequences		Third largest family			Second largest family			Largest family		
Number of monomers in chain	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

A useful measure of ruggedness is the autocorrelation function [16], 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 non-lethal 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:

$$R(t,s) = \frac{E(f_t \times f_{t+s}) - E(f_t) \times E(f_{t+s})}{variance(f)} \tag{4.1}$$

This process is repeated 100 times for each landscape and the mean autocorrelations for a selection of landscapes are plotted in Figure 4.2.

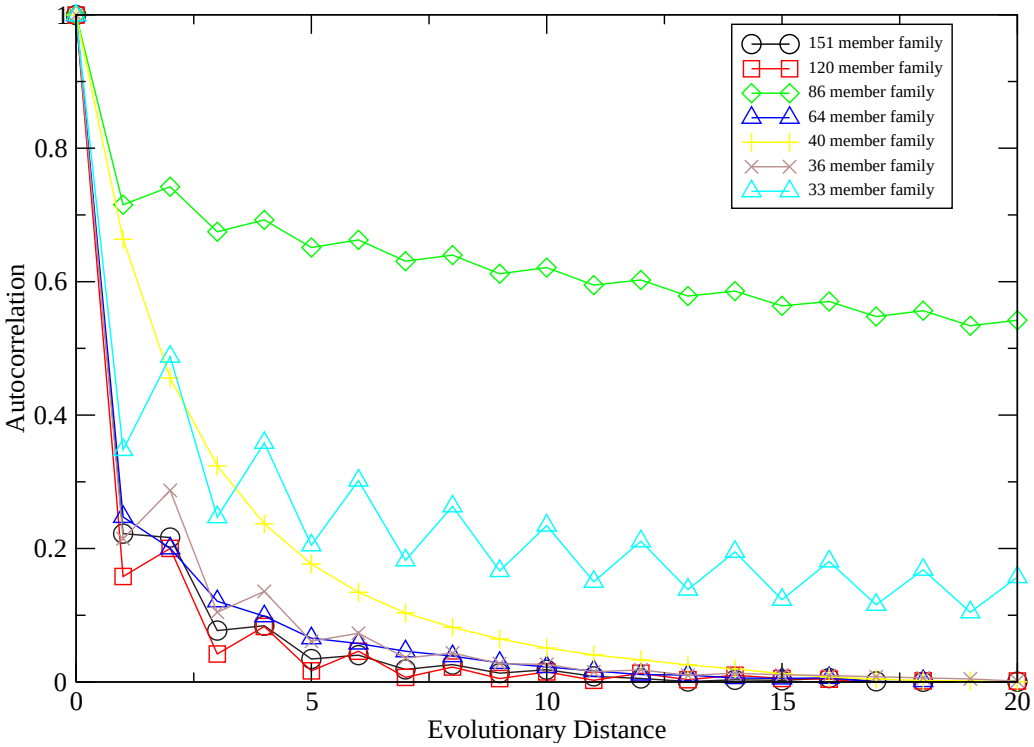


Figure 4.2: Autocorrelation for the evolutionary landscapes of various families.

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 behaviour, 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 residue, 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 behaviour, as the walk rarely crosses between the partitions.

The highly correlated nature of the 86-member family is due to the presence of two sub-landscapes, each of which exhibits a restricted range of fitnesses. Each sub-landscape 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 depicted as a graph in Figure 4.3. There are only two pathways that cross the partition, which we suggest is the cause of the high autocorrelation. These landscapes illustrate the variety of landscape structures that can be exhibited by minimalist models.

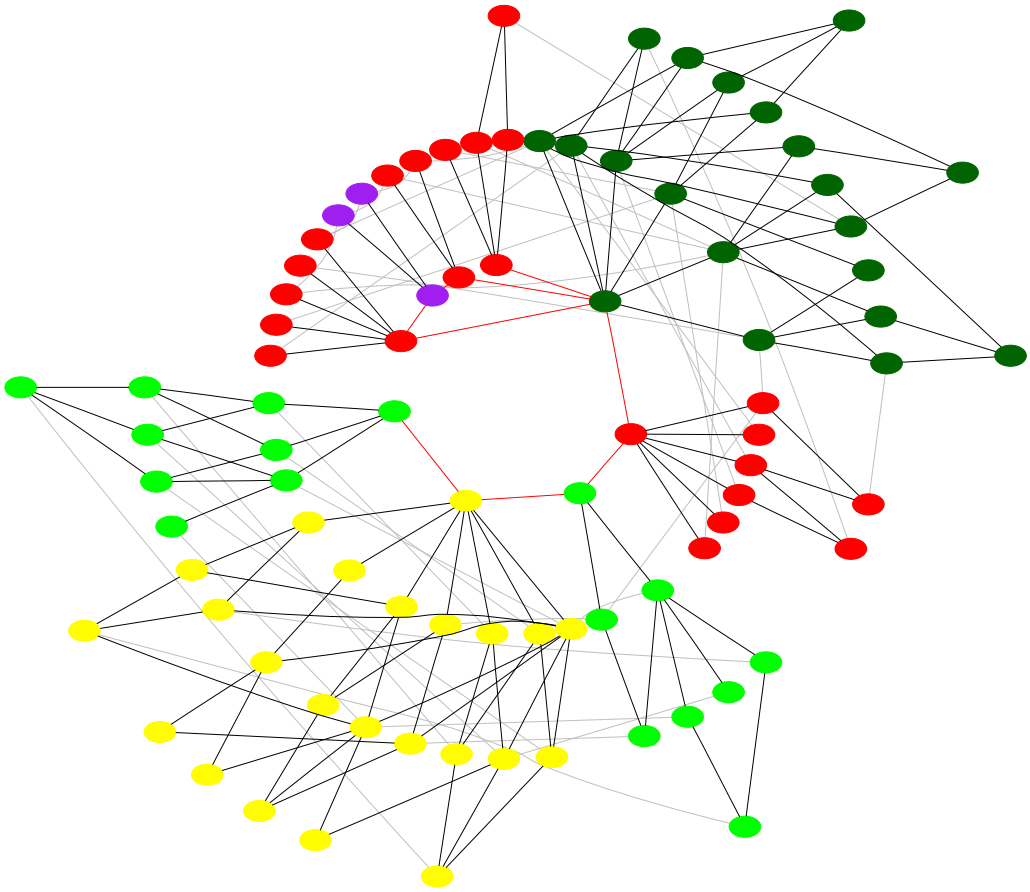


Figure 4.3: 86-member family of 25-mers. Nodes correspond to functional model proteins; edges correspond to single point mutations connecting two viable sequences. Hub nodes are closest to the centre, connected by red edges. Nodes are colour coded based on function (the number of H residues in the binding pocket, and the layer surrounding the binding pocket): yellow-3, light green-4, dark green-5, red-6, purple-7.

Table 4.3 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 4.3: Normalised 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

4.4 Visualisation 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 4.4. Examples of evolutionary landscapes are shown in Figure 4.3 and Figure 4.4. 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 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 visualise the landscapes by representing a viable sequence as a node coloured 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 4.4 depicts the largest landscape of 25-mers found, with 151 mem-

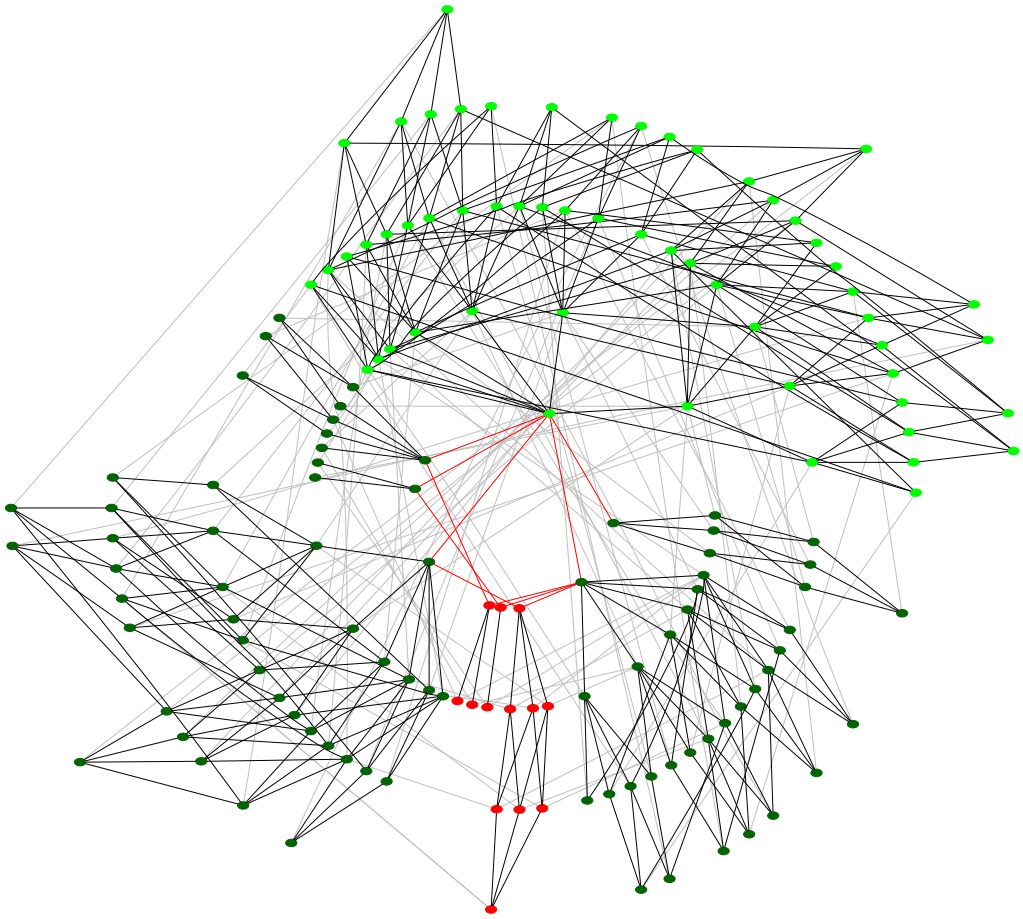


Figure 4.4: 151-member family of 25-mers, drawn with the same convention as Figure 4.3.

Table 4.4: Largest families for each chain length

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

bers. 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 then those with less. Thus, the areas of highest fitness are least able to withstand the rigours of mutation. The longer chains are able to support more hydrophobes more easily (Table 4.3), 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 Figure 4.1.

A feature discussed previously is the presence of critical edges [1], mutations that connect two otherwise unconnected portions of the landscape. These 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 Figure 4.3, where two smaller landscapes are linked through two mutations. These mutations link two different sub-landscapes, 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 [32], 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 12 families of 25-mers with 30 or more sequences and three or more neutral networks. 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.

4.5 Superfunnel structure of evolutionary landscapes

The presence of hub sequences on our landscapes is a manifestation of the superfunnel structure identified by Bornberg-Bauer and Chan [3]. They noted that, for their HP model, many landscapes were structured such that a prototype sequence, defined as that which had the most non-lethal mutations, was also the most stable sequence. The prototype sequence is surrounded by a layer of its one-mutant neighbours, which are all lower in stability than the prototype sequence. Outside this layer is the layer of two-mutant neighbours, which are lower in stability than the one-mutant neighbours. In this way each successive mutation away from the prototype sequence results in a decrease in stability.

We show how this structure coincides with our hub node structure in Figure 4.5. The stability of each sequence is calculated from the partition function. The probability that a sequence is in structure i with energy ε_i is given by

$$p_i = \frac{e^{\frac{-\varepsilon_i}{kT}}}{Q} \tag{4.2}$$

where Q is the partition function:

$$Q = \sum_i e^{\frac{-\varepsilon_i}{kT}} \tag{4.3}$$

The calculations are performed with $kT = 0.6$, as in reference [33], where k is the Boltzmann constant and T is the absolute temperature, although in the context of minimalist models kT may be considered as a single variable. In Figure 4.5 edges are directed, with the arrowhead pointing to increased p_n, the probability of being in the native structure, calculated from equation 4.2. However, unlike in reference [3], we plot the hub nodes at the centre, rather than a single prototype sequence. It can be seen that one of the hub nodes is also the prototype sequence; in this case the hub node sequence with four hydrophobes in the binding pocket and 12 allowed mutations. The implications of this structure of evolutionary landscapes on molecular evolution is discussed in chapter 5.

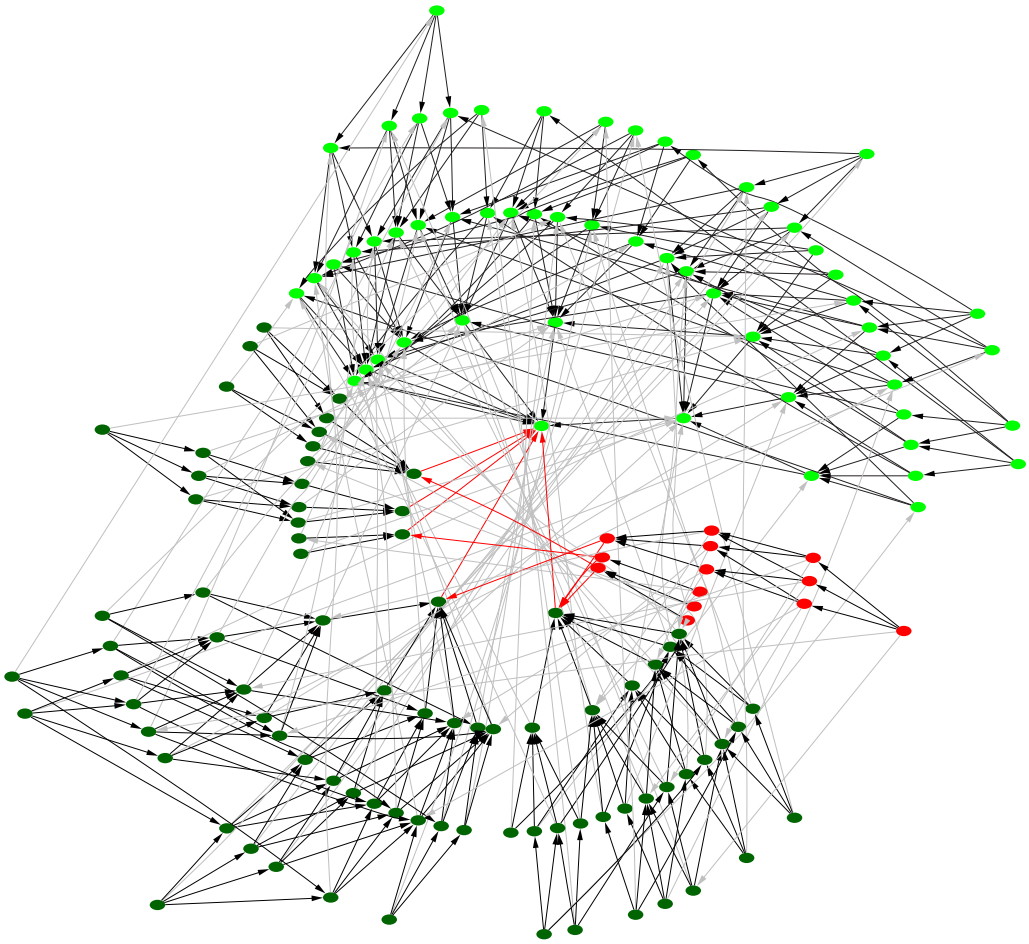


Figure 4.5: 151-member landscape, drawn as a superfunnel structure. The convention of Figure 4.3 is followed; however each edge is given a direction, indicated by the arrow, pointing towards increased stability.

4.6 Structural Mutations and Designability

In families previously examined on the square lattice we frequently observed mutations that altered the structure, whilst maintaining or altering function. A different behaviour 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 (the number of sequences folding to a particular structure) 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 non-bonded pairwise contacts. The Hamming distance is the number of differences between two contact maps. 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, firstly whether a structure can be a foldable protein structure at all, and secondly 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 to a highly designable structure. To consider this, we plot the average number of neighbours in contact map space in Figure 4.6 for a variety of 25-mer structures. 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 "protein-like" by ensuring the number of contacts per map is distributed identically to the 268 folding structures. We also plot the mean number of neighbours for only the more designable sequences. On Figure 4.6 the most designable structures are clearly more isolated from their neighbours than less designable structures. Also, the random structures have many more near neighbours 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 protein-like features such as secondary structure and tertiary symmetries such as β-barrels, four-helix bundles and α-helical coiled-coils

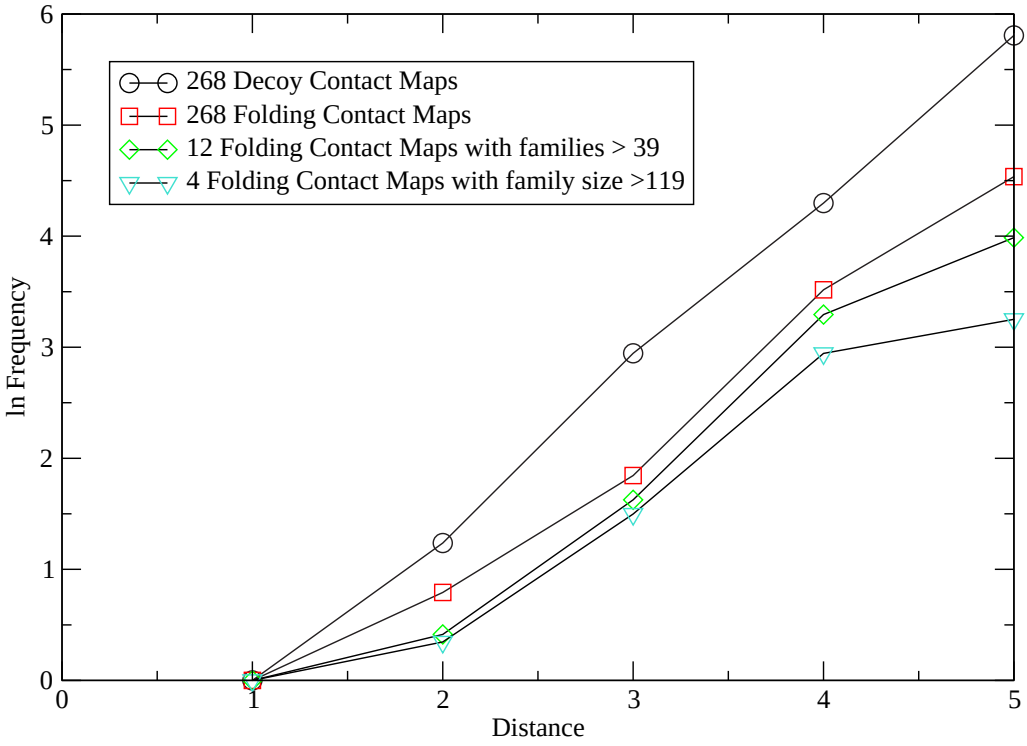


Figure 4.6: Contact map space. The logarithm of the mean frequency of occurrence of structures separated from a set of structures by a given Hamming distance, for several sets of sequences.

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 [31]. 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. Figure 4.7 shows the distributions of designability on the two lattices. The primary difference is that the diamond lattice exhibits a much smaller number of structures of low designability. This corresponds to the large families where structure is highly pliable. Figure 4.7(c) shows that this behaviour is not a simple consequence of the number of functional sequences found. The figure shows the 23-mer on the diamond 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 behaviour, 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

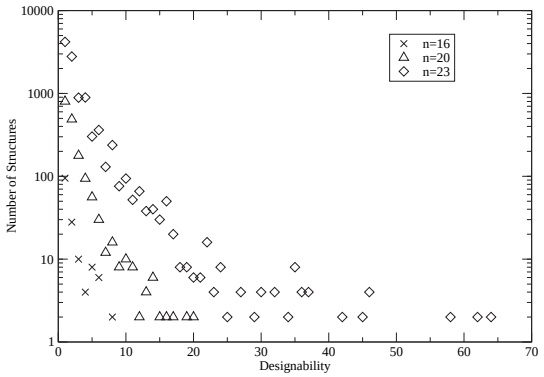
lattice [13, 25, 26]. We suspect that for the square lattice there are a large number of structures that are somewhat isolated in contact map space, and so are foldable, but poorly designable. On the diamond lattice, this may be different; the structures may be either highly isolated or well surrounded, with the intermediate case more of a rarity.

4.7 Conclusions

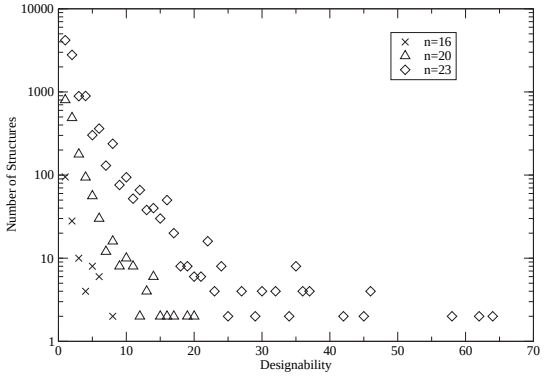
In this chapter, we have extended previous work [14, 1] by the study of functional model proteins on a 3D lattice. We have mutated shorter sequences in order to examine families of longer sequences. The requirement for function is one which makes functional model proteins rarer in our evolutionary landscapes than they would otherwise be. However, we believe it leads to a more accurate characterisation 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 the light of the evidence that the degree of structural mutations is tightly coupled to the choice of lattice.

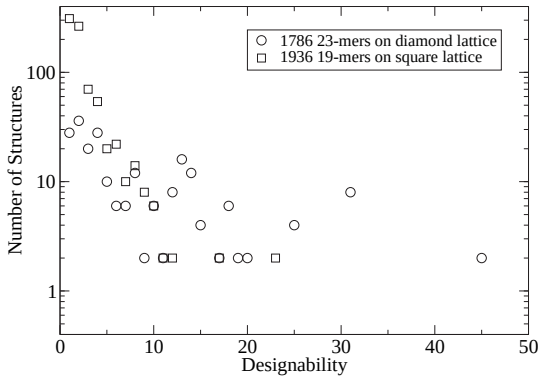
Designable structures were characterised as those which have highly unusual contact maps. Crippen [8] applied an empirical function to predict the number of structures with a given contact map on the cubic lattice. Examples of large deviations from this function were found, i.e. contact maps which



(a)



(b)



(c)

Figure 4.7: 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. Designability is the number of sequences that adopt a given structure.

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 [1] (page 70) 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 (Figure 4.3). This family consists of two structures organised into two sub- families connected by two edges. This is qualitatively similar to the effect of known mutations of Arc protein [7]. 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. Whilst 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.

Bibliography

[1] B.P. Blackburne and J.D. Hirst. Evolution of functional model proteins. *J. Chem. Phys.*, 115:1935–1942, 2001.

[2] E. Bornberg-Bauer. Randomness, structural uniqueness, modularity and neutral evolution in sequence space of model proteins. *Z. Phys. Chemie-Int. J. Res. Phys. Chem. Chem. Phys.*, 216:139–154, 2002.

[3] E. Bornberg-Bauer and H.S. Chan. Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. *Proc. Natl. Acad. Sci. U. S. A.*, 96:10689–10694, 1999.

[4] N.E.G. Buchler and R.A. Goldstein. Universal correlation between energy gap and foldability for the random energy model and lattice proteins. *J. Chem. Phys.*, 111:6599–6609, 1999.

[5] H.S. Chan and K.A. Dill. Comparing folding codes for proteins and polymers. *Proteins*, 24:335–344, 1996.

[6] J.J. Chou and E.I. Shakhnovich. A study on local-global cooperativity in protein collapse. *J. Phys. Chem. B*, 103:2535–2542, 1999.

[7] M.H.J. Cordes, N.P. Walsh, C.J. McKnight, and R.T. Sauer. Evolution of a protein fold in vitro. *Science*, 284:325–327, 1999.

[8] G.M. Crippen. Enumeration of cubic lattice walks by contact class. *J. Chem. Phys.*, 112:11065–11068, 2000.

[9] K.A. Dill. Dominant forces in protein folding. *Biochemistry*, 29:7133–7155, 1990.

[10] K.A. Dill, S. Bromberg, K.Z. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, and H.S. Chan. Principles of protein-folding - a perspective from simple exact models. *Protein Sci.*, 4:561–602, 1995.

[11] M. Eigen. Selforganization of matter and the evolution of macromolecules. *Naturwissenschaften*, 10:465–523, 1971.

[12] M.H. Hao and H.A. Scheraga. How optimization of potential functions affects protein folding. *Proc. Natl. Acad. Sci. U. S. A.*, 93:4984–4989, 1996.

[13] M.H. Hao and H.A. Scheraga. Molecular mechanisms for cooperative folding of proteins. *J. Mol. Biol.*, 277:973–983, 1998.

[14] J.D. Hirst. The evolutionary landscape of functional model proteins. *Protein Eng.*, 12:721–726, 1999.

[15] A. Irbäck and C. Troein. Enumerating designing sequences in the HP model. *J. Biol. Phys.*, 28:1–15, 2002.

[16] S. Kauffmann. *The Origins of Order: Self-Organization and Selection in Evolution.* Oxford University Press, Oxford, 1993.

[17] W. Kauzmann. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.*, 14:1–63, 1959.

[18] H. Kaya and H.S. Chan. Towards a consistent modeling of protein thermodynamic and kinetic cooperativity: How applicable is the transition state picture to folding and unfolding? *J. Mol. Biol.*, 315:899–909, 2002.

[19] M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, 1983.

[20] H. Li, R. Helling, C. Tang, and N. Wingreen. Emergence of preferred structures in a simple model of protein folding. *Science*, 273:666–669, 1996.

[21] L. Li, L.A. Mirny, and E.I. Shakhnovich. Kinetics, thermodynamics and evolution of non-native interactions in a protein folding nucleus. *Nat. Struct. Biol.*, 7:336–342, 2000.

[22] Y. Mandel-Gutfreund and L.M. Gregoret. On the significance of alternating patterns of polar and non- polar residues in beta-strands. *J. Mol. Biol.*, 323:453–461, 2002.

[23] S.A. Marshall and S.L. Mayo. Achieving stability and conformational specificity in designed proteins via binary patterning. *J. Mol. Biol.*, 305:619–631, 2001.

[24] D.A. Moffet and M.H. Hecht. De novo proteins from combinatorial libraries. *Chem. Rev.*, 101:3191–3203, 2001.

[25] E. Shakhnovich and A. Gutin. Enumeration of all compact conformations of copolymers with random sequence of links. *J. Chem. Phys.*, 93:5967–5971, 1990.

[26] E.I. Shakhnovich. Protein design: a perspective from simple tractable models. *Fold. Des.*, 3:R45–R58, 1998.

[27] H. Taketomi, Y. Ueda, and N. Gō. Studies on protein folding, unfolding and fluctuations by computer simulation. i. the effect of specific amino acid sequence represeneted by specific inter-unit interactions. *Int. J. Peptide Protein Res.*, 7:445–459, 1975.

[28] D.M. Taverna and R.A. Goldstein. Why are proteins so robust to site mutations? *J. Mol. Biol.*, 315:479–484, 2002.

[29] G. Tiana, R.A. Broglia, and D. Provasi. Designability of lattice model heteropolymers. *Phys. Rev. E*, 64:art. no.–011904, 2001.

[30] F. T. Wall and F. T. Hioe. The distribution of end-to-end lengths of self-avoiding walks on the diamond lattice. *J. Chem. Phys.*, 74:4410–4415, 1970.

[31] T.R. Wang, J. Miller, N.S. Wingreen, C. Tang, and K.A. Dill. Symmetry and designability for lattice protein models. *J. Chem. Phys.*, 113:8329–8336, 2000.

[32] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[33] P.D. Williams, D.D. Pollock, and R.A. Goldstein. Evolution of functionality in lattice proteins. *J. Mol. Graph.*, 19:150–156, 2001.

[34] S. Wright. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the Sixth International Congress on Genetics*, pages 354–366, 1932.

[35] T. Yomo, S. Saito, and M. Sasai. Gradual development of protein-like global structures through functional selection. *Nat. Struct. Biol.*, 6:743–746, 1999.

[36] K. Yue, K.M. Fiebig, P.D. Thomas, H.S. Chan, E.I. Shakhnovich, and K.A. Dill. A test of lattice protein-folding algorithms. *Proc. Natl. Acad. Sci. U. S. A.*, 92:325–329, 1995.

Chapter 5

Population dynamics simulations of functional model proteins

5.1 Introduction

The study of the evolutionary landscapes detailed in the previous chapters can be extended by running computer-based simulations of the movement of populations on the landscape. We employ a deterministic population dynamics simulation. An assumption of the model is that the population is infinite and so, given that the landscape does not contain any disconnected regions, simulations will always converge to the same point. The advantage of this is that it allows us to measure the progress of the simulation in terms of a known end point. The disadvantage is that certain aspects of the landscape, such as possible "kinetic traps", cannot be fully analysed.

Population dynamics simulations depend on knowledge of a fitness landscape. This landscape must describe sequence space, the allowed transitions between sequences (i.e. which sequences are neighbours in sequence space), and the mapping from sequence space to a phenotype. The NK model [8] and the "block NK" model [14, 1] provide such treatments of evolutionary landscapes. The NK model offers a tunably rugged landscape through its treatment of epistatic interactions. Each of the N elements represents a gene, or an amino acid in a gene sequence, depending on the chosen for of the model. Each element is given a different, randomly assigned, fitness that depends on the genotype of K other elements. This dependence of a gene's fitness on the rest of the genome is known as epistasis. The fitness is given by the mean of the fitnesses of the N elements. At low levels of epistasis ($K = 0$) the landscape forms a single peak. For higher values of K the landscape increases in ruggedness, till $K = N - 1$, where the landscape is uncorrelated.

One problem with the NK landscape is that, for large values of N, the central limit theorem applies, and so a mutation is likely to have little effect on fitness. The block NK model addresses this by partitioning each sequence into a number of blocks. This replaces the epistatic effect of K. A mutation in a block has an effect only on the fitness contribution of other sites in the block. Unlike the NK model, where K is constant over all N, blocks can vary in size. The fitness is defined as the sum of the fitness of each block.

The NK model and its derivatives offer tunable landscapes that allow the effect of landscape structure on evolution to be investigated. Another approach that addresses realistic biological landscapes is to model the known

features of a biological system, such as the affinity maturation of the immune response. The binding of an antigen to a B cell will cause it to begin to reproduce by cloning. Mutant descendants of these B cells may produce antibodies that bind the antigen more strongly, and its proliferation will be be up-regulated. In this way, somatic hypermutation of B cells results in the production of antibodies with increased affinity. A theoretical treatment of this system was developed [9], and later used to study the effect of variable mutation rates on evolution [10]. The model demonstrated an advantage to periods of high mutation rate followed by periods of stable growth over a static mutation rate.

The effects of the requirements of protein folding and function can be directly addressed by modelling protein structure explicitly with molecular dynamics simulations. Using such a model, Yomo *et al.* [19] evolved protein sequences to adopt a certain binding pocket conformation. They found that compactness, helical content, and folding ability all result from this selection for function. However, simulations at atomic resolution are computationally expensive. For this reason lattice models offer an attractive method for generating a genotype-phenotype mapping, as they are both physically relevant and computationally tractable [4]. science A population dynamics simulation of lattice model proteins gave results consistent with the molecular dynamics simulation of Yomo *et al.* [19]; evolution of a lattice model protein for compactness and evolution for function do not result in greatly different populations [16].

The effect of the allowed transitions in sequence space has been tested with a lattice model for genotype-phenotype mapping [3]. The introduction of

recombination as an allowed mutation permitted existing diversity to be quickly amplified, and on some occasions otherwise inaccessible boundaries in sequence space to be crossed. However, point mutations were still essential for effective exploration of sequence space. Another study examined a similar model [18], and finds that a restricted ratio of recombination to point mutations is necessary to find an optimal sequence on a flat landscape.

Here we employ a definition of landscape that is limited to those sequences that are minimally fit (i.e. fold to a stable structure and possess a binding pocket). This means that the whole of sequence space for length $n = 25$, for example, will consist of a set of disconnected landscapes. A lethal mutation is considered as stepping off the landscape by this definition, and non-fit sequences are either not drawn in our landscape graphs, or labelled explicitly.

5.2 Methods

5.2.1 Modelling Evolution

We adapted a model used to examine evolution with recombination [3]. The progress of evolution is governed by equation 5.1. This relates the population $p_i(q+1)$ of a viable sequence i in generation $q+1$ to the populations of generation q.

$$p_i(q+1) = \mathcal{N}(q) \left\{ (1-\mu)\, p_i(q) + \frac{\mu}{n} \sum_{j=1}^{n} p_j(q) \right\} f_i \qquad (5.1)$$

Here, μ is the point mutation rate, equivalent to μ_m in reference [3]. It is defined as the fraction of population that mutates each generation. The population of sequence i at generation q is $p_i(q)$. Thus, $\mu p_i(q)$ is the fraction of population lost to mutations at generation q. The length of the chain is n. Since we study a two-letter alphabet there are n sequences that differ from i by a single point mutation. Each sequence, j, of these n sequences donates $\frac{\mu}{n} p_j(q)$ of population to sequence i. The factor $\mathcal{N}(q)$ normalises the populations so that:

$$\sum_i p_i(q+1) = 1$$

Each generation, the effect of fitness is accounted for by multiplying the population by a factor, f_i, the fitness of sequence i. In reference [3] it is a function of stability. Here we define it similarly, but as a function of the fitness defined earlier, the number of hydrophobes around the binding pocket.

$$f_i = e^{\alpha h_i} \tag{5.2}$$

Here, we incorporate both the fitness, as defined by the number of hydrophobes in the binding pocket (h_i), and a selection coefficient (α), which can be considered to be analogous to an inverse temperature. At low α, evolution becomes more neutral; at high α, selection will dominate when possible.

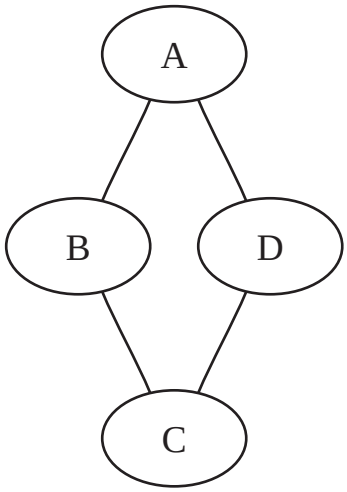


Figure 5.1: A toy evolutionary landscape. Each node represents a sequence, and each edge an allowed mutation.

5.2.2 Implementation

The simulations can be treated as a set of matrix operations on a population vector. We begin with a treatment of this system as a Markov chain process [6]. Markov chains describe the random transition of a system between discrete states. Strictly speaking they apply to the prediction of the probability of finding the system in a certain state after a given number of transitions. However, this probability is equivalent to the proportion of an infinite population existing in the given state.

We first consider the toy network in figure 5.1. This is the landscape for HP sequences of length two (i.e. HH HP PH and PP). Ignoring fitness, and assuming sequence space is constrained to those four sequences (i.e. no lethal mutations) then we can construct a Markov chain for population dynamics on this landscape. We begin by writing the connectivity matrix

for this landscape, which simply lists all allowed mutations on the landscape.

$$C = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \tag{5.3}$$

From this matrix we can derive a transition matrix, M, for the Markov process. This gives the probability of mutation from state i to state j ($\frac{\mu}{2}$, as $n = 2$). The leading diagonal consists of the probability of no mutation $(1 - \mu)$:

$$M = \begin{pmatrix} 1-\mu & \frac{\mu}{2} & 0 & \frac{\mu}{2} \\ \frac{\mu}{2} & 1-\mu & \frac{\mu}{2} & 0 \\ 0 & \frac{\mu}{2} & 1-\mu & \frac{\mu}{2} \\ \frac{\mu}{2} & 0 & \frac{\mu}{2} & 1-\mu \end{pmatrix} \tag{5.4}$$

We define a row population vector, $\mathbf{p}(q)$ for generation q:

$$\mathbf{p}(q) = \begin{pmatrix} p_A(q) & p_B(q) & p_C(q) & p_D(q) \end{pmatrix} \tag{5.5}$$

The next generation, $\mathbf{p}(q+1)$ can be obtained from $\mathbf{p}(q)$ and M:

$$\mathbf{p}(q+1) = \mathbf{p}(q)M \tag{5.6}$$

The treatment of this process as a Markov chain allows us to find the steady

state of the populations. There exists some a for which M^a has all non-zero components (for $\mu > 0$). This means that M is a regular probability matrix, and so converges to a limit matrix, $T = \lim_{a\to\infty} M^a$. From knowledge of T, we can find the steady state population, given by the fixed probability vector $t = \mathbf{p}(0)T$. In this case, M converges to:

$$T = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} \tag{5.7}$$

and so the steady state is given by:

$$t = \mathbf{p}(0)T = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} \tag{5.8}$$

This treatment does not allow for the possibility of lethal mutations; our population cannot move off the landscape. To extend this model to capture the process detailed in equation 5.1, we must include a treatment of lethal mutations, fitness, and the renormalisation factor. To illustrate this we define a new toy landscape that includes lethal mutations (for $n > 2$) in figure 5.2.

We begin by redefining our population vector, $\mathbf{p}$, to take into account lethal mutations. The vector for the general case is given below:

$$\mathbf{p}(q) = \begin{pmatrix} p_1(q) & p_2(q) & p_3(q) & \dots & p_i(q) & p_d(q) \end{pmatrix} \tag{5.9}$$

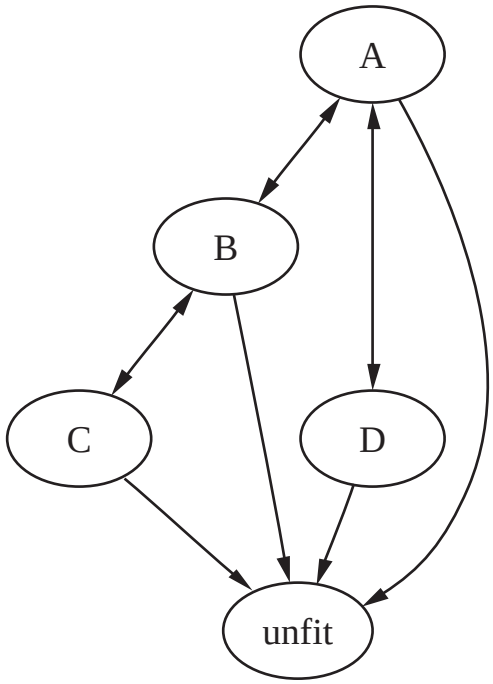


Figure 5.2: A toy evolutionary landscape. Nodes A,B,C and D represent viable sequences. Mutations to lethal sequences are represented by the node labelled as “unfit”. Edges are directed, indicating that mutations to unfit sequences are allowed, but unfit sequences do not input population to viable sequences.

The population lost to lethal mutations (redistributed during renormalisation) is $p_d(q)$. The population of each viable sequence, i, in the network being examined is $p_i(q)$.

We can redefine the connectivity matrix C and transition matrix M from figure 5.2.

$$C = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{5.10}$$

$$M = \begin{pmatrix} 1-\mu & \frac{\mu}{n} & 0 & \frac{\mu}{n} & \frac{(n-2)\mu}{n} \\ \frac{\mu}{n} & 1-\mu & \frac{\mu}{n} & 0 & \frac{(n-2)\mu}{n} \\ 0 & \frac{\mu}{n} & 1-\mu & 0 & \frac{(n-1)\mu}{n} \\ \frac{\mu}{n} & 0 & 0 & 1-\mu & \frac{(n-1)\mu}{n} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{5.11}$$

In this second toy network, there are n possible mutations at each sequence, of which, some are lethal. Lethal mutations are represented by the node labelled as "unfit" in figure 5.2. In order to accommodate these possible mutations, an extra column and row are added in equations 5.10 and 5.11. The final column of M is the probability of a lethal mutation from that sequence. Due to the connectivity of the landscape some sequences will undergo more lethal mutations than others. The final row of M is the set of transition probabilities for non-viable sequences. In this simulation non-

viable sequences never mutate spontaneously to viable sequences as they are unable to exist and reproduce, so the probability that they stay unfit is 1.

The process as given is now an absorbing Markov chain [6]. Population dynamics will proceed until the whole population is in the absorbing state. In this case the absorbing state is the set of unfit sequences. There are two more variables from equation 5.1 that have yet to be incorporated: the fitness, f_i of each sequence, and the normalising factor $\mathcal{N}$. These two quantities prevent the system from moving to the absorbing state. We introduce a row vector, $\mathbf{f}$, that contains the fitness f_i of each sequence i. The final entry in this vector gives the fitness of the unfit sequences, defined as zero. The effect of fitness on each generation is given by $\mathbf{p}(q) \diamond \mathbf{f}$, i.e. the Hadamard, or entry-wise, product:

$$\mathbf{p}(q)M = \begin{pmatrix} p_1'(q) & p_2'(q) & p_3'(q) & \dots & p_i'(q) & p_d'(q) \end{pmatrix} \tag{5.12}$$

$$\mathbf{f} = \begin{pmatrix} f_1 & f_2 & f_3 & \dots & f_i(q) & 0 \end{pmatrix} \tag{5.13}$$

$$(\mathbf{p}(q)M) \diamond \mathbf{f} = \begin{pmatrix} f_1 p_1'(q) & f_2 p_2'(q) & f_3 p_3'(q) & \dots & f_i(q) p_i'(q) & 0 \end{pmatrix} \tag{5.14}$$

Finally, multiplication by $\mathcal{N}_q$ renormalises the population.

$$\mathbf{p}(q+1) = \mathcal{N}_q((\mathbf{p}(q)M) \diamond \mathbf{f}) \tag{5.15}$$

The renormalisation and fitness factors mean that the process is no longer a Markov chain and the steady state cannot be found from the convergence of the transition matrix. We begin our population dynamics experiments by "seeding" each simulation with a population of unity at a certain sequence. The simulation proceeds until it has converged to its steady state. We then repeat this process for each sequence in the family. We examine the largest 25-mer family of 151 sequences, the second largest family of 120 sequences and the 86 sequence family, which is notable for the presence of a bottleneck and kinetic trap.

5.3 Finding the steady state

A practical effect of the effectively infinite population in our deterministic population dynamics simulation is that steady state is never truly reached, instead evolution proceeds asymptotically towards the steady state. We monitor the rate of evolution using the sum of the squares of the differences in population from generation to generation, counted for each of the C sequences in the landscape examined.

$$\delta_p(q) = \sum_{i=1}^{C} (p_i(q) - p_i(q-1))^2 \tag{5.16}$$

Figure 5.3 plots $\ln \delta_p$ for a population dynamics simulation beginning at one particular sequence (HPPHPPPHPPPPPHHPHPHPPPPHH) in the 86-member family of 25-mers with $\alpha = 0.3$, $\mu = 0.01$. The trend we observe for δ_p is an exponential decrease, until the values of p_i do not differ to the

precision of a double precision floating point number.

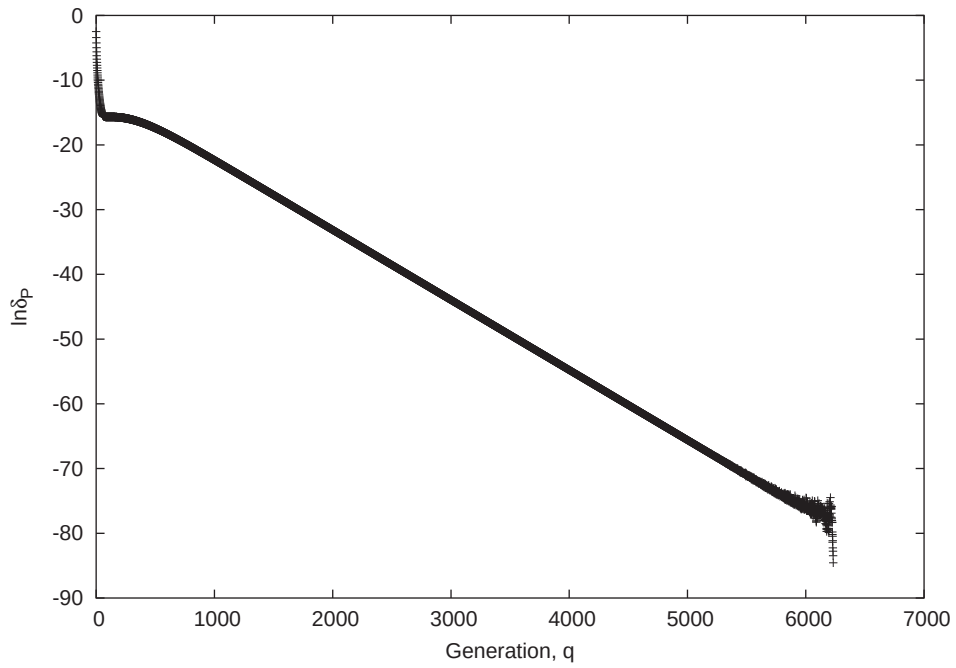


Figure 5.3: Rate of population movement over the course of a population dynamics experiment.

In order to assess the end point of a population dynamics experiment we choose a cut-off level, terminating the simulation when $\delta_p < 10^{-20}$. Fluctuations below this level are insignificant, unless we are concerned with very small changes in large ($>10^{10}$) populations of organisms.

5.4 Threshold Selection

Initially, we consider how populations are distributed on our landscapes under fit/not-fit selection. This is also known as threshold selection, where

all sequences with fitness greater then zero are considered equally fit. The knowledge of how populations are distributed under threshold selection will help in understanding the distribution of populations when a more realistic selection pressure is applied.

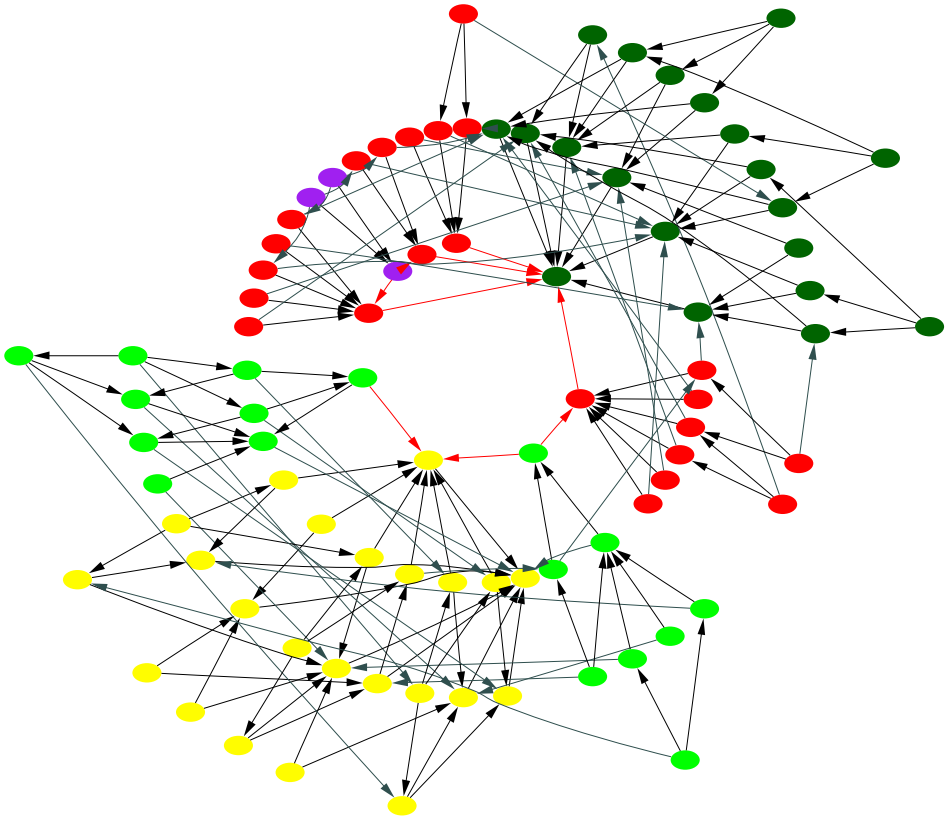


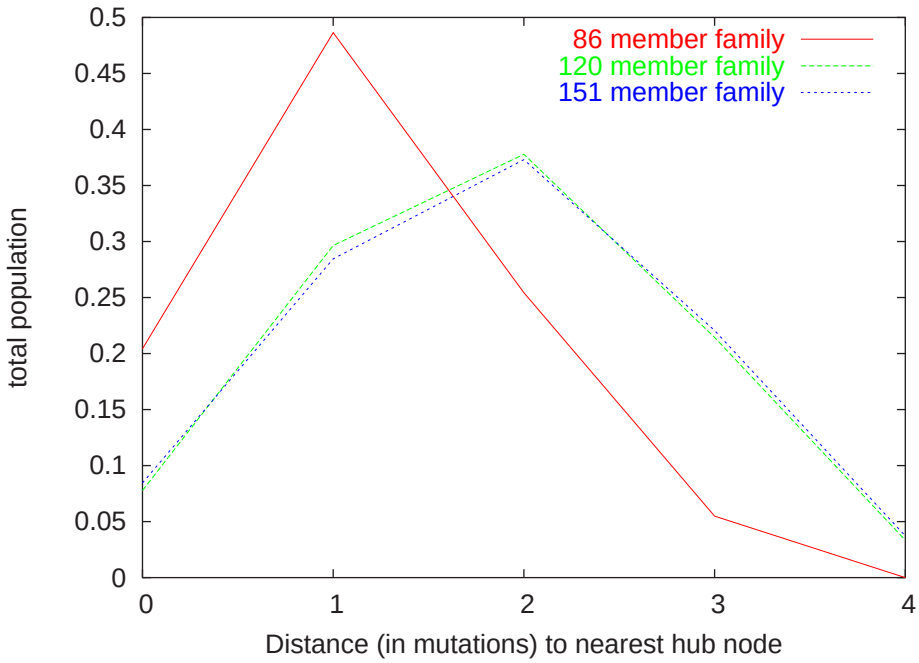
Figure 5.4: 86-member family of 25-mers. Nodes correspond to functional model proteins; edges correspond to single point mutations connecting two viable sequences. Nodes are colour coded based on function (the number of H residues in the binding pocket): yellow-3, light green-4, dark green-5, red-6, purple-7. Red edges are used to indicate connections between hub nodes. Arrow direction indicates increasing population.

Figure 5.4 shows how the populations vary over one particular landscape at

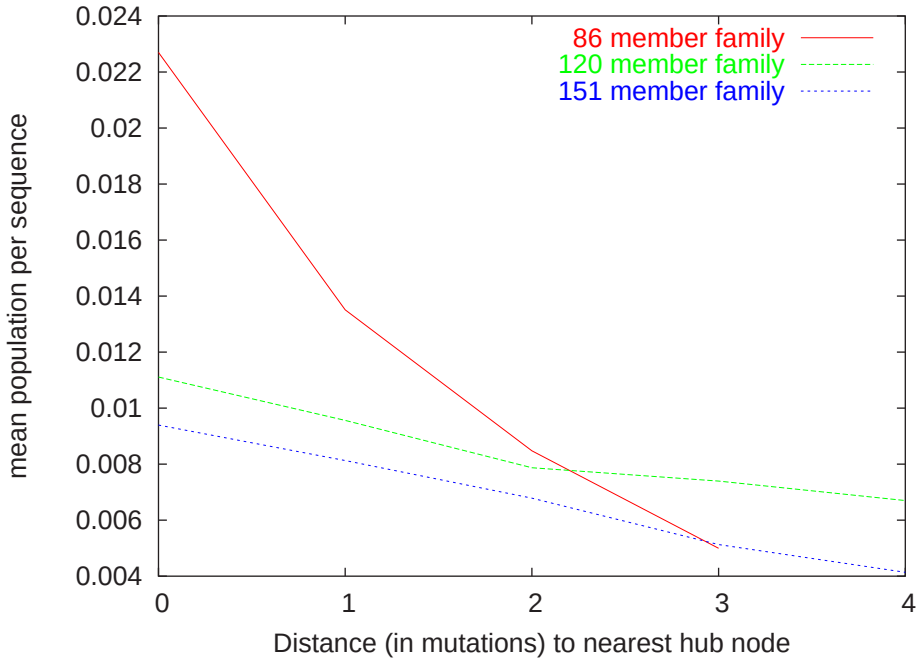
the steady state under threshold selection. As might be expected, the hub nodes are among the most populous sequences. However, the bulk of the population still occupies non-hub nodes, due to the weight of numbers. Taverna and Goldstein [17] reason that the nature of evolutionary landscapes is such that the number of highly stable sequences is greatly outweighed by the number of marginally stable sequences. Their population dynamics simulations demonstrate that this is a sufficient cause for the marginal stability of real proteins. In the language of our landscapes, this means that in practice the sequence fixed during the course of evolution will most likely be a node near the rim of the superfunnel.

Taverna and Goldstein note that for the majority of stable sequences most mutations will lower the thermodynamic stability. We see from figure 5.4 why this should be - the majority of mutations for a sequence lead away from the hub and so are climbing "up the superfunnel". This reduction in thermodynamic stability does not necessarily correspond with a reduction in fitness.

Our results are broadly in agreement with these observations. Figure 5.5(a) shows how total population varies with distance from the hub for three of the larger networks, under threshold selection, with $\mu = 0.01$. The larger families have their population distributed further away from the hub than the smaller family. However, the connectivity of the landscape complicates the distribution of sequences. Although the central, most thermodynamically stable sequences may be heavily outnumbered by the less stable sequences, the connectivity of the landscape causes the more thermodynamically stable sequences to be over-represented. This is shown in figure 5.5(b). The



(a) Distribution of population under threshold selection



(b) Distribution of mean population per sequence under threshold selection

Figure 5.5: Distance from the hub of (a) total population and (b) mean population per sequence at steady state under a population dynamics simulation.

sequences close to the hub have a higher population per sequence than those further from the hub. This is consistent with the results of Bornberg-Bauer and Chan [2], and Xia and Levitt [18]. In the event of the fixation of a single sequence, or a small number of sequences, this will likely result in the fixation of sequences of marginal thermodynamic stability away from the hub. However in the case of evolution under high mutation rates (e.g. viruses) then the over-representation of the more stable sequences means that they will occur more frequently than simple consideration of the proportion of stable sequences would suggest. Furthermore, recombination may result in larger populations at the centre of the landscape [18].

5.4.1 Connection between population and connectivity

In figure 5.6 we plot the number of nearest neighbours against the population for two landscapes. In both these cases there appears to be some weak correlation. Greater correlation is evident when plotting the number of next-nearest neighbours (Figure 5.7). The 151-member landscape shows a high level of correlation ($R^2 = 0.97$). However, the 86-member landscape shows two different groups, both with their own trend. This means that one group has a higher population than would be predicted from the correlation, and one group a lower population. The colours of the points in figure 5.7(a) refer to the fitness of the corresponding node, and match the colour scheme in figure 5.4. The two different groups on figure 5.7(a) correspond to the two "sub-landscapes" on figure 5.4, that are linked by only two allowed mutations. The two points that lie between the two obvious groupings in figure 5.7(a) correspond to the two nodes that lie closest to the other sub-

landscape.

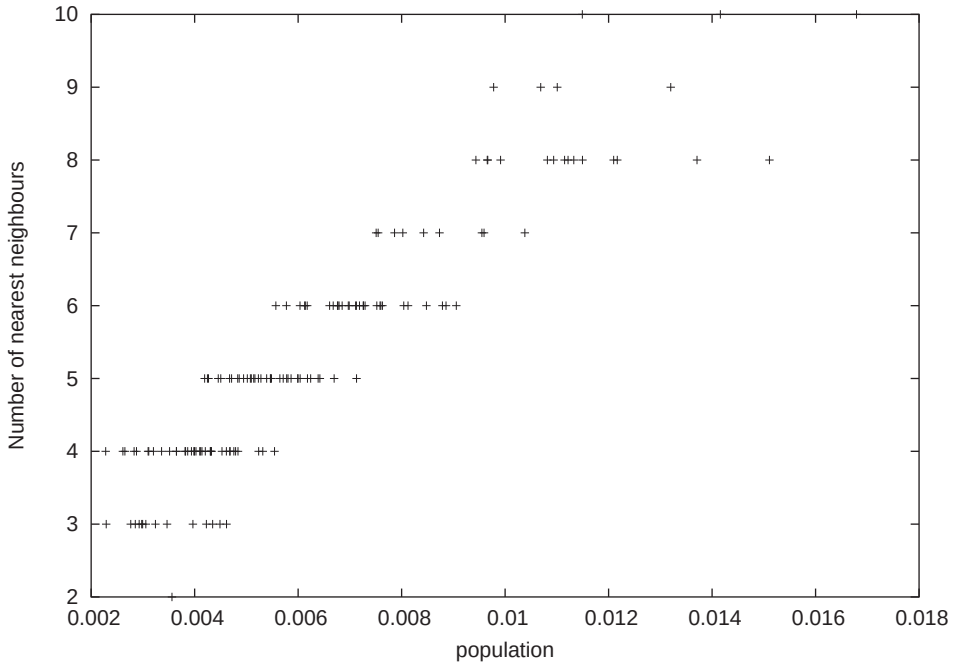
5.5 “Entropy” and “enthalpy”

5.5.1 Effect of α and μ

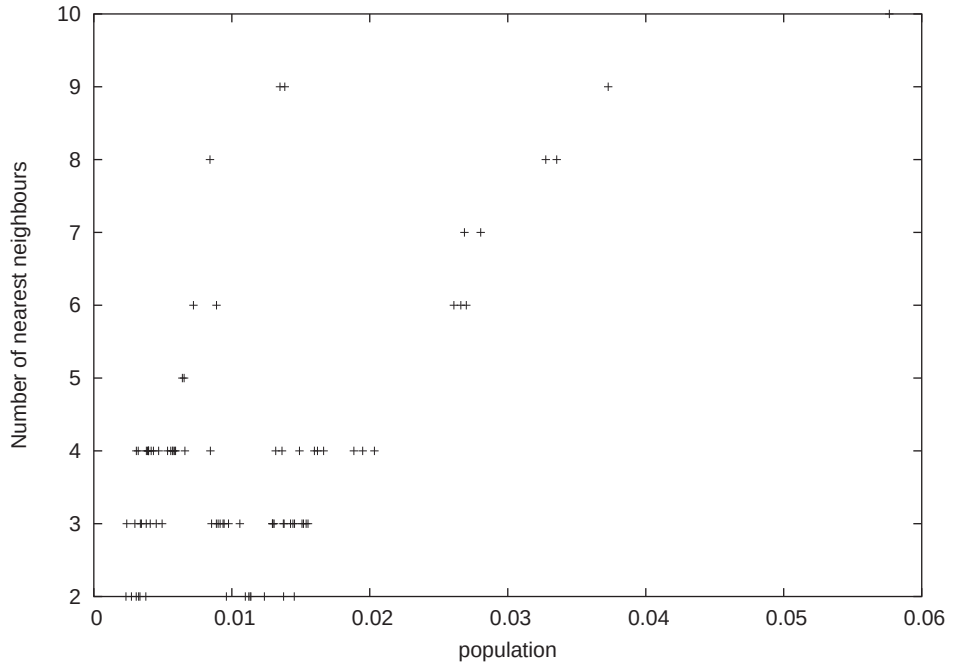
In order to gain some insight into the principles that govern the populations in our simulations we consider population dynamics on a toy landscape. Figure 5.8 shows another simplified landscape, with directional edges given to show explicitly the proportion of mutants that move in each direction. The landscape is for sequences of length $n = 3$, allowing three possible mutations. We consider the effect of varying α and μ on this landscape, after setting the fitness of A and A$'$ as two, and B as one. This is plotted in figure 5.9. Due to the topology of the landscape, the populations of A and A$'$ will always be equal in the steady state population.

The selection coefficient, α, acts as an inverse temperature. This is demonstrated in figure 5.9. At “low temperature”, where α is high, the population is fixed at nodes of class A (A and A$'$). Although at each generation A and A$'$ mutate to node B, at high α the fitness of B is too low and this population dies before reproduction. Figure 5.10 shows the reverse situation, where B is more fit than A and A$'$. In this case, the population of B rises to 1 at high temperatures.

The effect of low α (“high temperature”) is to move the system to a state governed by the effects of a kind of “entropy”. We use the term entropy

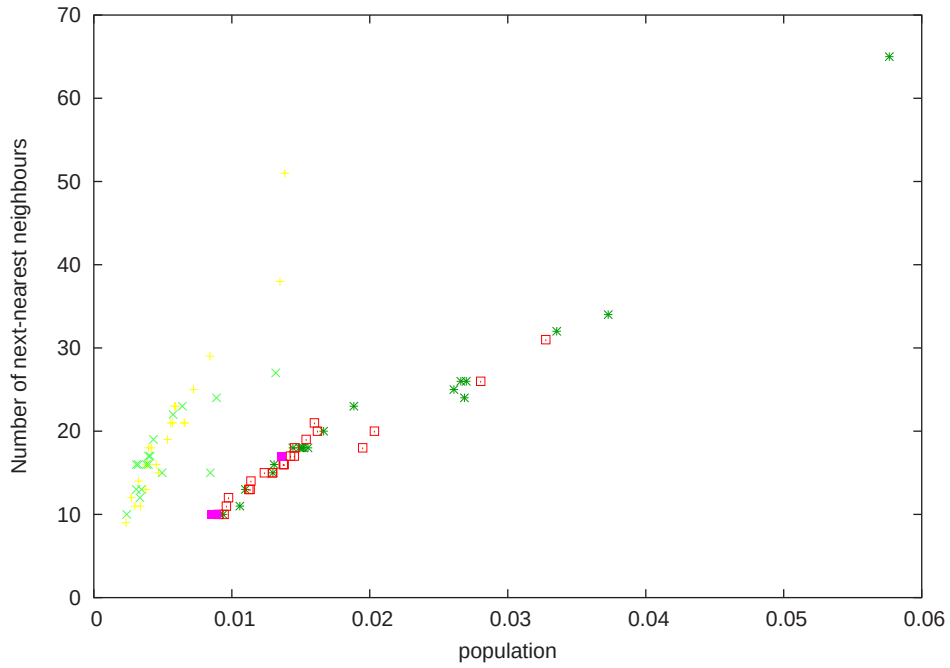


(a)

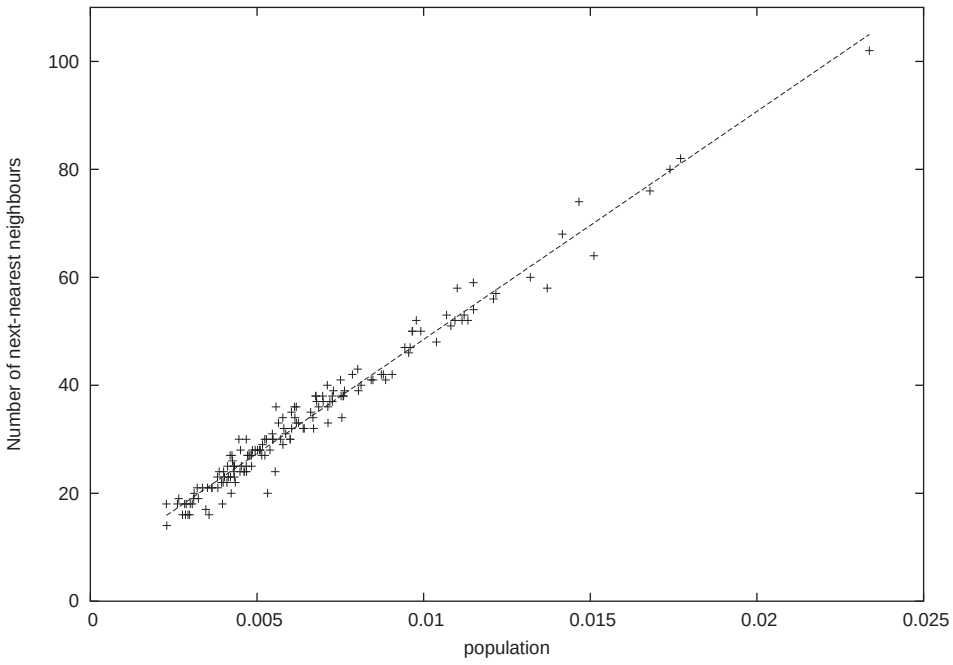


(b)

Figure 5.6: The relationship between the steady-state population and the number of nearest neighbours under threshold selection for (a) the 86-member landscape and (b) the 151-member landscape.



(a)



(b)

Figure 5.7: The relationship between the steady-state population and the number of next-nearest neighbours under threshold selection for (a) the 86-member landscape and (b) the 151-member landscape. Colour in (a) shows the fitness of the sequence according to the scheme in figure 5.4

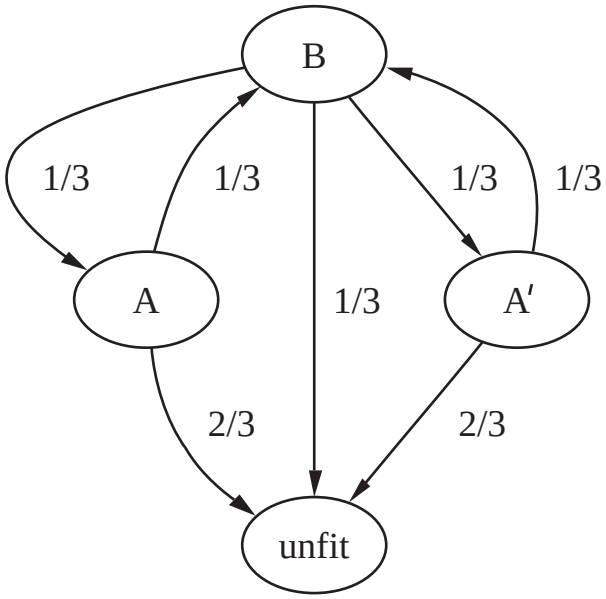


Figure 5.8: A toy evolutionary landscape. Nodes A, A′ correspond to the three viable sequences on the landscape. Lethal mutations are explicitly accounted for by the node labelled "unfit". Edges are directed - once a sequence has mutated to a non-viable sequence, it cannot reproduce and mutate back. The sequences are given the length $n = 3$. The values attributed to each edge gives the proportion of mutations in that direction. For A and A′, $2/3$ of the mutations will be lethal, for B, only $1/3$.

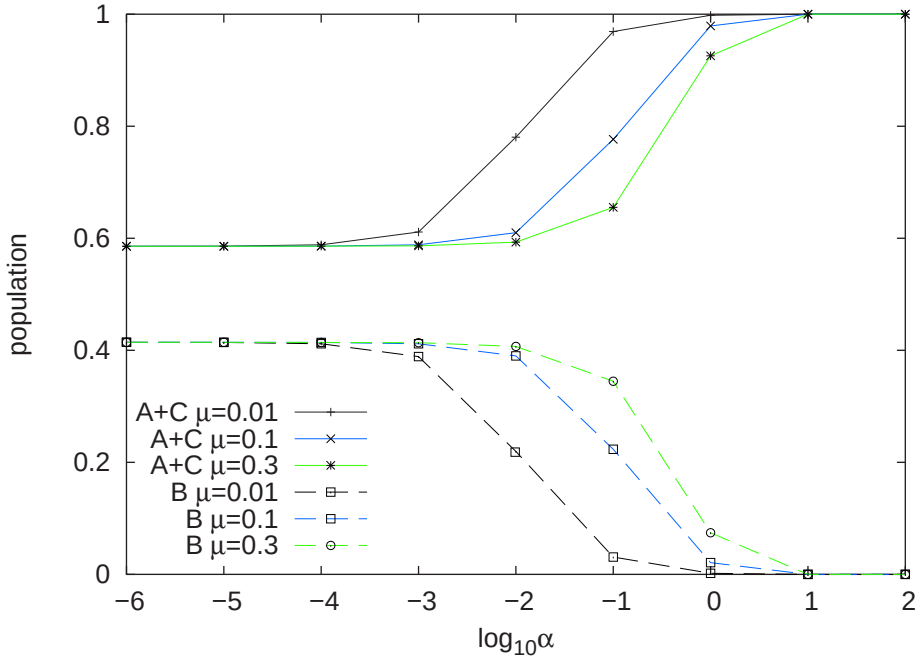


Figure 5.9: Population dynamics simulations on the toy landscape given in figure 5.8. The fitness of node B is set at 1, A and C are set to 2.

here to draw an analogy with energetic landscapes, however entropy (and enthalpy) on these landscapes cannot be directly connected to free energy, as the process of population dynamics is not identical to the process of chemical kinetics. For this reason, we refer to “pseudo-entropy” and “pseudo-enthalpy” when referring to processes on the landscape. Pseudo-entropy is different to a sum of the number of nodes - it also involves the connectivity of the graph. Hence the population of node B is not $1/3$ when $\alpha = 0$, but, after the equation governing the landscape has been solved, can be shown to be $\frac{1}{1+\sqrt{2}} \approx 0.414$ (see below).

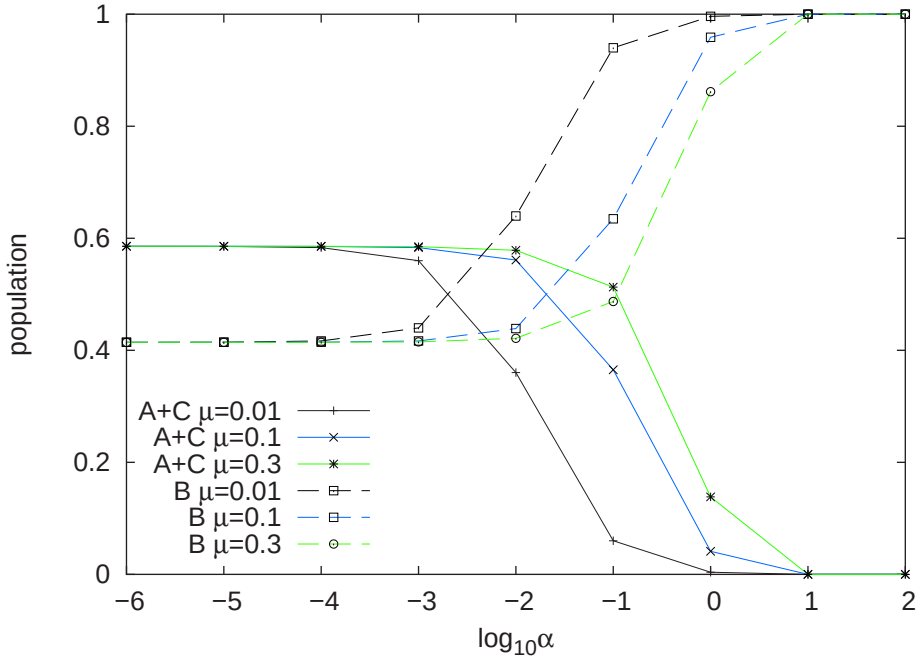


Figure 5.10: Population dynamics simulations on the toy landscape given in figure 5.8. The fitness of node B is set at 2, A and C are set to 1.

5.5.2 Solving the toy landscape

For simple landscapes, the equation that determines the steady state population can be solved. For a system with one node of type B and one or more nodes of type A we can calculate the population from one generation to the next:

$$a_{q+1} = \mathcal{N}(q)\left((1-\mu)a_q + \frac{\mu b_q}{n}\right)e^{\alpha h_a} \tag{5.17}$$

$$b_{q+1} = \mathcal{N}(q)\left((1-\mu)b_q + c_a\frac{\mu a_q}{n}\right)e^{\alpha h_b} \tag{5.18}$$

For simplicity, the above equation deals with A nodes as equal in popula-

tions. At the steady state, those nodes will always be equal in population, and the above equations hold if the population dynamics process is begun with equal populations on each A node. In the above equations, a_q and b_q are the populations of individual nodes of class A and B at generation q, h_a and h_b are the fitnesses of nodes of class A and B respectively and n is the length of the sequence. The number of nodes of class A is given by c_a ($c_a = 2$ in Figure 5.8). The normalisation factor for each generation is $\mathcal{N}(q)$.

At the steady state, $\frac{a_q}{b_q}$ is constant from generation to generation. Hence, for populations at the steady state denoted by a_s and b_s, we can combine equations 5.17 and 5.18:

$$\frac{a_s}{b_s} = \frac{\left((1-\mu)a_s + \frac{\mu b_s}{n}\right) e^{\alpha h_a}}{\left((1-\mu)b_s + c_a\frac{\mu a_s}{n}\right) e^{\alpha h_b}} \tag{5.19}$$

The populations of A and B are constrained to sum to 1:

$$c_a a_q + b_q = 1 \tag{5.20}$$

We can solve these two equations to find roots for a and b. The equation of the solution is long; we give the solution for when $\alpha = 0$ (i.e. when all fitnesses are considered equal) below:

$$a_s = \frac{1}{\sqrt{c_a} + c_a} \tag{5.21}$$

$$b_s = \frac{1}{1+\sqrt{c_a}} \tag{5.22}$$

When $\alpha = 0$, μ, n, f_a and f_b have no effect. When $\alpha > 0$ these factors do have an effect. To illustrate the non-trivial effect of these factors, and for completeness, we give the solution for a_s:

$$a_s = \frac{-\left(e^{h_b\,\alpha}\,n\,(-1+\mu)\right) - e^{h_a\,\alpha}\,(n-(2\,c_a+n)\,\mu)}{-2\,c_a\,e^{h_b\,\alpha}\,(n\,(-1+\mu)+\mu)+2\,c_a\,e^{h_a\,\alpha}\,(n\,(-1+\mu)+c_a\,\mu)} \tag{5.23}$$
$$-\frac{\sqrt{\left(e^{2\,h_a\,\alpha}+e^{2\,h_b\,\alpha}\right)\,n^2\,(-1+\mu)^2 - 2\,e^{(h_a+h_b)\,\alpha}\,\left(n^2\,(-1+\mu)^2 - 2\,c_a\,\mu^2\right)}}{-2\,c_a\,e^{h_b\,\alpha}\,(n\,(-1+\mu)+\mu)+2\,c_a\,e^{h_a\,\alpha}\,(n\,(-1+\mu)+c_a\,\mu)}$$

5.6 "Thermodynamics"

If we consider h_a and h_b as pseudo-enthalpic factors, then we consider α, μ and n to be components of a "temperature". By determining a pseudo-entropy for a node or a neutral net of nodes, we can attempt to predict steady state populations under different conditions for our networks. The complexity of the solution for even a simplified toy landscape (equation 5.23) suggests that any thermodynamic approach will be an approximation. How well this approximation fits could be a useful measure of the simplicity of a landscape.

To explore this concept, we propose that the partition function can be used to estimate populations on the landscape. Clearly we can imagine land-

scapes that would display behaviour that is not easily predictable from the partition function. Consider two evolutionary landscapes (figure 5.11). One is a node of high fitness surrounded by an area of low fitness. A second is identical, but the low fitness area is higher in fitness. The pseudo-entropy of each node will be identical, since whatever definition of pseudo-entropy we choose for our landscapes, it must be independent of the pseudo-enthalpy (fitness), and in all respects besides fitness these networks are identical. The central nodes in each network will therefore possess equal pseudo-entropy and pseudo-enthalpy but will have different populations.

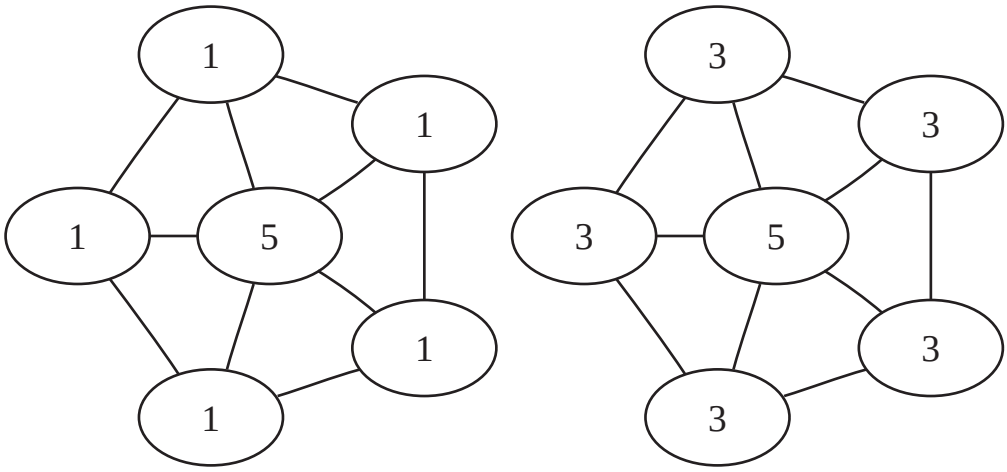


Figure 5.11: Two possible networks that illustrate aspects of pseudo-entropy and pseudo-enthalpy. Each node represents a sequence, while edges represent allowed point mutations. Numbers indicate the fitness of a node.

This means that a simple calculation based on our definitions of pseudo-entropy and pseudo-enthalpy cannot always capture the steady state populations of these landscapes. Nevertheless, it may be that the landscapes formed by our functional model proteins are not as extreme in their structure. If we observe that, in practice, populations on the landscapes of func-

tional model proteins do behave in a manner broadly akin to chemical thermodynamics, this tells us about the nature of the landscapes.

5.6.1 Use of the partition function to estimate populations

The partition function, Q, gives the population, p_i, associated with an energy, ε_i, and a factor β.

$$p_i = \frac{e^{-\beta\varepsilon_i}}{Q} \tag{5.24}$$

where:

$$Q = \sum_j e^{-\beta\varepsilon_j} \tag{5.25}$$

In the case where $\beta = 0$, all states will be equally populated. This is clearly not the case with our population dynamics simulations (figure 5.6) where, due to the connectivity of the landscape there will be different occupancies even when $\alpha = 0$ (i.e. when $\beta = 0$). This is analogous to a thermodynamic landscape, where the effects of entropy alter the populations of different energy levels according to their degeneracies. For a set of w_i degenerate states of i, the total population of those states will be given by:

$$p_i = w_i \frac{e^{-\beta\varepsilon_i}}{Q} \tag{5.26}$$

We accommodate the effect of connectivity on the landscape by treating each sequence as a set of degenerate states in the partition function. We define w_i as a function, σ of the number of next nearest neighbours for sequence i, λ_i:

$$p_i = \sigma(\lambda)\frac{e^{-\beta\varepsilon_i}}{Q} = \frac{e^{-\beta\varepsilon_i + \ln\sigma(\lambda_i)}}{Q} \tag{5.27}$$

Thus, if we can ind an expression to predict an pseudo-enthalpy ε_i, pseudo-entropy $\ln\sigma(\lambda)$, and β (which on a thermodynamic landscape is equivalent to $1/kT$), we can estimate the population of a sequence or a neutral net of sequences.

5.6.2 Determination of pseudo-enthalpy and pseudo-entropy

We define the pseudo-enthalpy of a sequence (or set of sequences) from the fitness of the set of sequences:

$$\varepsilon_i = -h_i \tag{5.28}$$

There is no direct relationship between the number of states of a molecule and the connectivity of the landscape around a sequence, which would lead to a strict definition of pseudo-entropy. A relationship between physical systems and our landscapes can be drawn, however, from consideration of the population at high "temperature". Under high temperature conditions, where enthalpy becomes a negligible consideration, the entropy dominates

the free energy equation. In the same way, we can consider entropy of a sequence to be related to its steady-state population when $\alpha = 0$, i.e. under threshold selection.

The method for determining pseudo-entropy is based on the observation that there is a high correlation between the number of next-nearest neighbours and the population of sequences at steady-state under threshold selection. This was considered in section 5.4.1. This means we are defining pseudo-entropy from the local structure of the landscape; for this reason we refer to it as "local entropy". The presence of this correlation can be explained from the nature of the population dynamics simulation. During the simulation each sequence loses the same amount of population as every other sequence through mutations. For threshold selection the only way to gain population is through renormalisation (which cannot increase the population of one sequence over another), or through receiving donated population from a neighbour. Thus there is some correlation between the number of nearest neighbours and population under threshold selection. However, the amount of population donated by the neighbours depends on their own populations, which in turn depends on their own neighbours. This correlation shows that a good approximation to this effect is a count of next-nearest neighbours. On larger landscapes, it may be advantageous to extend this to further reaches of the landscape.

A sequence that has two paths to another sequence donates an equivalent amount of population as two sequences connected by a single path each. Another way to understand this is that two neighbouring nodes have their populations boosted by this next-nearest neighbour, rather than just one.

Because of this, when we calculate the local entropy of a node, we count a next-nearest neighbour one time for each path. The data shown in figure 5.7 includes this "double counting".

The correlation between next-nearest neighbours and pseudo-entropy depends on the population of each node not deviating too sharply from the mean; i.e. all nodes are considered equal. More highly populated nodes will have more population to donate to their neighbours through mutation. If the populations vary widely for some other reason, for example during selection, or because of an unusual landscape structure, the results will be skewed away from the predictions. Because of this, the partition function is poor at estimating the population of individual sequences within a neutral net. Consider a population under high selective pressure (high α). At the steady state, the population will be almost entirely on the nodes of highest fitness. The local entropy is calculated from all next-nearest neighbours, even though only the nodes of highest fitness will have any population to donate. For this reason, the internal distribution of a population within a neutral net will not be predicted accurately from the partition function. Instead we apply it to the estimation of populations of all sequences of each fitness, calculating the local entropy as the total number of next-nearest neighbours for all sequences of a given fitness.

The relationship between population and number of next-nearest neighbours was found to be linearly well correlated when $\alpha = 0$ (Figure 5.7(b)), i.e.; of

the form:

$$p_i = \mathcal{K}\lambda_i + \mathcal{C} \tag{5.29}$$

Here, λ_i is the number of next-nearest neighbours for sequence i; $\mathcal{K}$ and $\mathcal{C}$ are constants, and the value of $\mathcal{C}$ is negligible. Furthermore, the majority of sequences have a large number of next-nearest neighbours. We can assume that $\mathcal{C} = 0$ for simplicity. For reasons already outlined, we wish to consider whole populations of sequences with identical fitnesses:

$$p_s = \mathcal{K}\sum_{i=0}^{c_s}\lambda_i + \mathcal{C} \tag{5.30}$$

Here, p_s refers to the total population of the c_s sequences of a given fitness. We know that $\sum_{j=0}^{c} p_j = 1$, for all c sequences. Hence, we can eliminate the constant $\mathcal{K}$, and $\mathcal{C}$ is zero:

$$p_s = \mathcal{K}\sum_{i=0}^{c_i}\lambda_i \tag{5.31}$$

$$\sum_{j=0}^{c} p_j = \mathcal{K}\sum_{j=0}^{c}\lambda_j \tag{5.32}$$

$$\mathcal{K} = \frac{1}{\sum_{j=0}^{c}\lambda_j} \tag{5.33}$$

$$p_s = \frac{\sum_{i=0}^{c_s}\lambda_i}{\sum_{j=0}^{c}\lambda_j} \tag{5.34}$$

From the partition function (equation 5.27):

$$\frac{\sum_{i=0}^{c_s} \lambda_i}{\sum_{j=0}^{c} \lambda_j} = \frac{\sigma(\lambda_s) e^{-\beta \varepsilon_s}}{Q_{\alpha=0}} \tag{5.35}$$

$Q_{\alpha=0}$ is the partition function when $\alpha = 0$, and is constant for a given landscape. Recall that $\alpha = 0$, and so $\beta = 0$:

$$\frac{\sum_{i=0}^{c_i} \lambda_i}{\sum_{j=0}^{c} \lambda_j} = \frac{\sigma(\lambda_s)}{Q_{\alpha=0}} \tag{5.36}$$

We define a reduced component of the local entropy found above as $\sigma^{'}(\lambda_s)$ that excludes the partition function $Q_{\alpha=0}$:

$$\sigma^{'}(\lambda_i) = \frac{\sum_{i=0}^{c_s} \lambda_s}{\sum_{j=0}^{c} \lambda_j} \tag{5.37}$$

Hence:

$$\ln \sigma(\lambda_s) = \ln \sigma^{'}(\lambda_s) + \ln Q_{\alpha=0} \tag{5.38}$$

5.6.3 Determining β, the "inverse temperature"

We return to the partition function:

$$p_s = \frac{e^{-\beta \varepsilon_s + \ln \sigma(\lambda_s)}}{Q} \tag{5.39}$$

As already stated, we wish to consider the populations of all sequences of a certain fitness, rather than individual sequences. Thus, the equation is a re-written version of equation 5.27; p_s is the predicted total population of all sequences of set s; ε_s is the enthalpy of sequences in set s. We combine equations 5.28, 5.38 and 5.39 below:

$$p_s = \frac{e^{h_s + \ln \sigma'(\lambda_s) + \ln Q_{\alpha=0}}}{\sum_z e^{h_z + \ln \sigma'(\lambda_z) + \ln Q_{\alpha=0}}} = \frac{\left(e^{h_s + \ln \sigma'(\lambda_s)}\right) Q_{\alpha=0}}{\sum_z \left(e^{h_z + \ln \sigma'(\lambda_z)}\right) Q_{\alpha=0}} \tag{5.40}$$

$$p_s = \frac{e^{h_s + \ln \sigma'(\lambda_s)}}{\sum_z e^{h_z + \ln \sigma'(\lambda_z)}} \tag{5.41}$$

Equation 5.41 defines the estimation of total population of each fitness from the known factors h_s (the fitness of the set of sequences s) and $\sigma'(\lambda_s)$ (equation 5.37), and the unknown factor β (the "inverse temperature"). The sum in the denominator counts for each set, z, of sequences of identical fitness. As demonstrated with the toy landscape, β must be a function of α, μ and n. We expect the effect of an increase in α to be an increase in β (a decrease in "temperature").

Clearly an increase in μ should increase the "temperature" (decrease β), by allowing sequences of lower fitness to still be populated by mutations from other sequences. Each generation, $\frac{\mu}{n}$ of the population is donated to each neighbouring sequence. Therefore the effect of n is to oppose the effect of μ; an increase in n will increase β. The effect of a small change in n is likely to be small, and since in this chapter we exclusively consider landscapes where $n = 25$, we neglect it as a factor.

We attempt to characterise the relationship between μ, α and β empirically. We vary α from 0.0005 to one, keeping μ constant at 0.1, and run a population dynamics simulation on the largest, 151-member landscape (Figure 5.12) until the steady state is reached. We find a value of β by solving equation 5.41 with p_s taken as the population of the set of sequences with fitness six. We repeat this process for values of μ from 0.001 to 0.9, keeping α constant at 0.1. We then fit functions to these data. The functions are given below, where β_α is the fit to α and β_μ the fit to μ. These functions are plotted with the data used to generate them in figure 5.13.

$$\beta_\alpha = 17.3535 + \frac{.0410204}{\alpha^{0.670229}} - \frac{10.1479}{\alpha^{0.107302}} \tag{5.42}$$

$$\beta_\mu = \frac{1}{0.126652 + 1.4264\mu - 3.841719\mu^2 + 5.37244\mu^3} \tag{5.43}$$

5.7 Results

5.7.1 Predictions of populations from the partition function

This thesis attempts to address the reasons why some landscapes are well fitted by the prediction from the partition function, rather than to derive a general function that is applicable in all cases. Therefore, for reasons of clarity, only variation of α is considered further; μ is kept fixed.

We use β_α to make predictions of the populations for several different land-

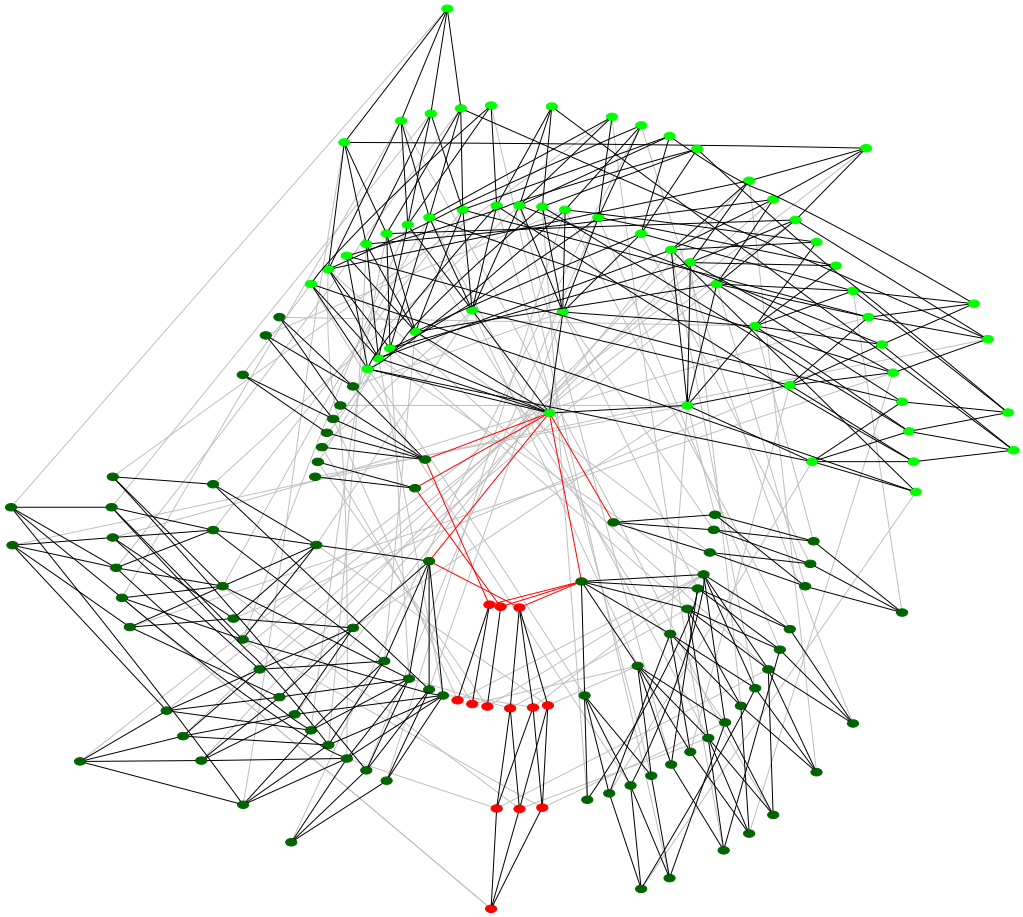
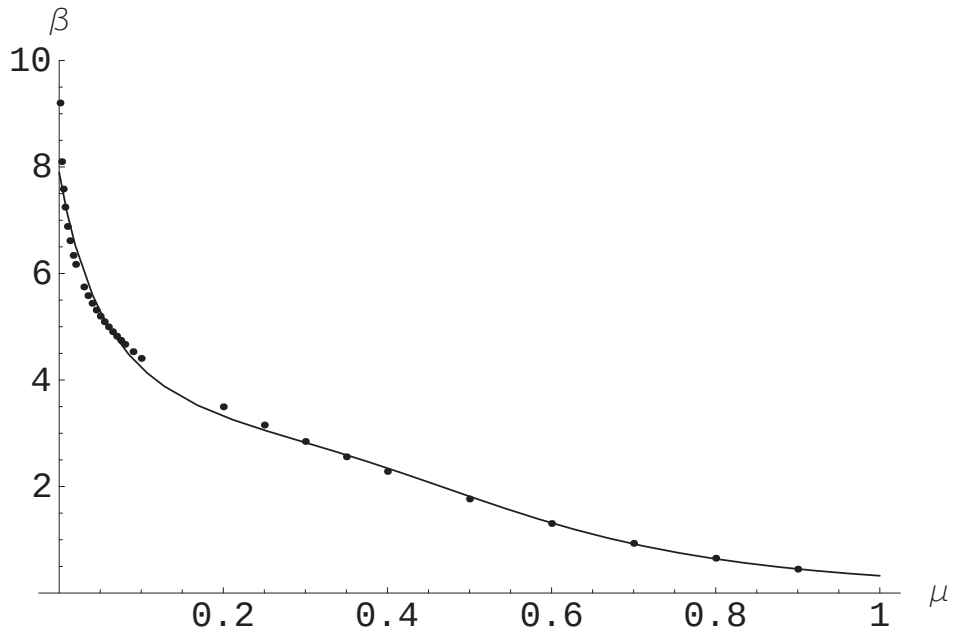
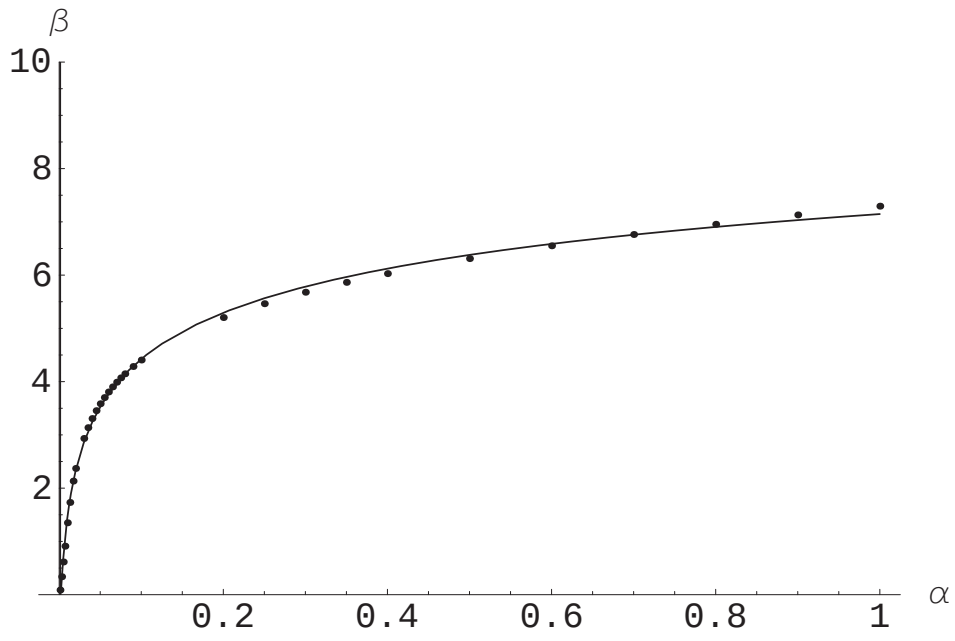


Figure 5.12: 151 member landscape of three dimensional 25-mer functional model proteins, drawn with the same convention as figure 4.3.



(a) The relationship between μ and β on the 151-member landscape for $\alpha = 0.1$



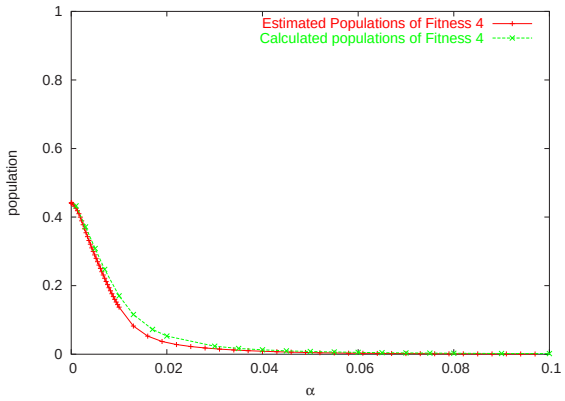
(b) The relationship between α and β on the 151-member landscape for $\mu = 0.1$

Figure 5.13: The relationship between β (i.e. an "inverse temperature" function) and (a)μ, (b)α. The functions are fitted with the equations detailed in the body of the text.

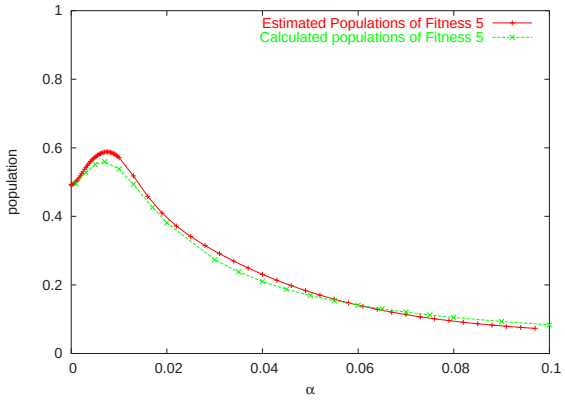
scapes under different conditions. Figure 5.14 shows the results for the 151-member landscape. Clearly the method is effective at estimating the populations in this case. This is to be expected, as the method used the values from this population of fitness 6 to generate β - effectively the estimation is a fit to the calculated values. However, it is reassuring that the method captures the trends quantitatively for the other two fitnesses which were not explicitly included in the development of the β function.

The next largest landscape has 120 sequences and is shown in figure 5.15. Again, the free-energy method makes good estimations of the populations (figure 5.16). A landscape with a much different structure is the 86-member landscape (figure 5.4). This has is structured as two sub-landscapes, as discussed in the previous chapter (page 100). This structure makes populations on this landscape difficult for the free-energy method to estimate. The population curves are qualitatively, although not quantitatively, captured (figure 5.17).

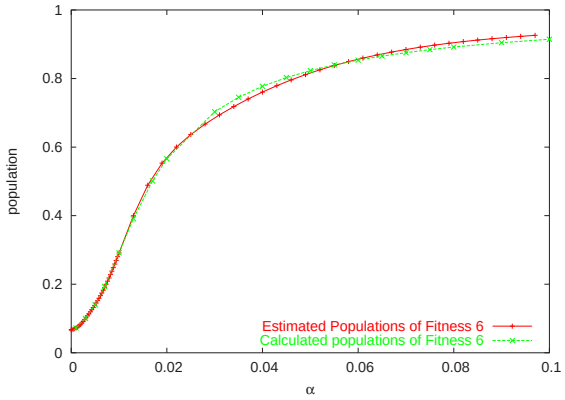
The reasons for the difference between estimated and real populations on the 86-member landscape can be found from the validity of the assumption described earlier, that the number of next nearest neighbours is a good indication of population at the steady-state under threshold selection. This is not the case for this landscape (see Figure 5.7(a)), which leads to the discrepancies seen in figure 5.17 at low α. At higher α these functions give a better estimate. We suggest that the inaccurate predictions at low α arise from that the structure of the landscape; there are effectively two landscapes in competition with each other. Consider another toy landscape, shown in figure 5.18. Two sub-landscapes have a weak interconnection, but are well



(a)



(b)



(c)

Figure 5.14: Populations of (a) fitness 4, (b) fitness 5 and (c) fitness 6 for the 151-member landscape (shown in figure 5.12). Populations are calculated through the population dynamics method and estimated using the free-energy method described in the text.

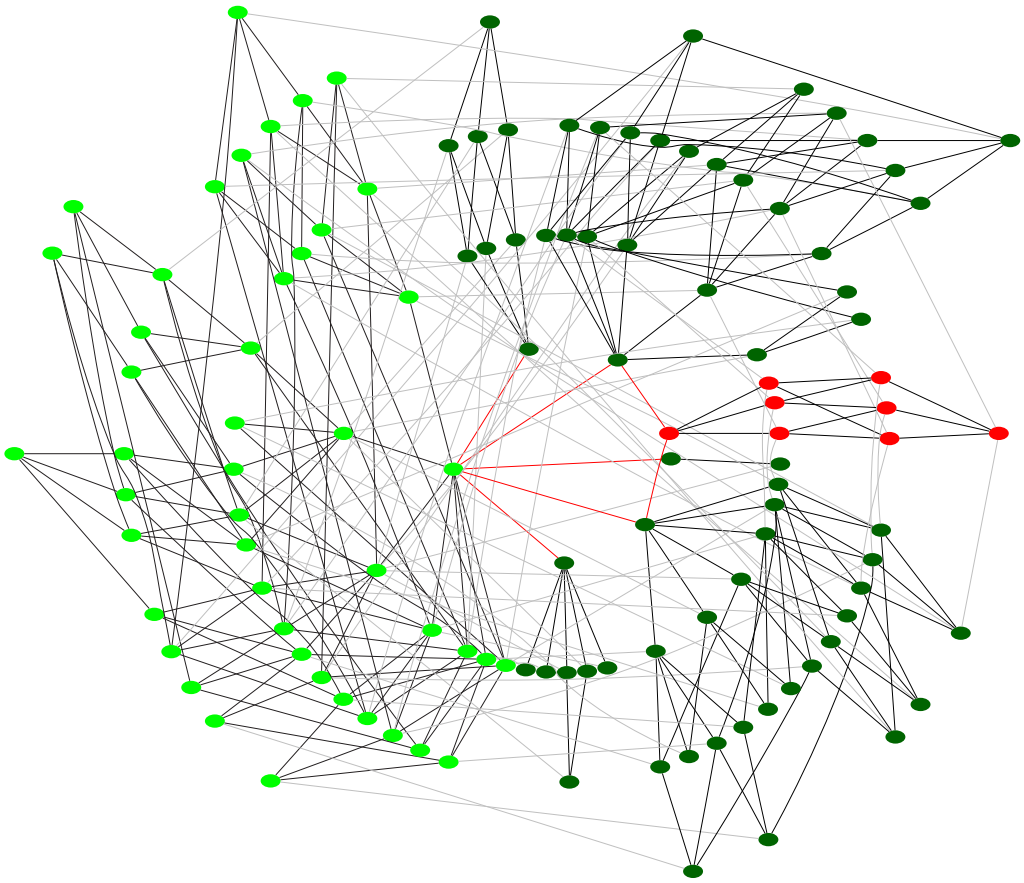
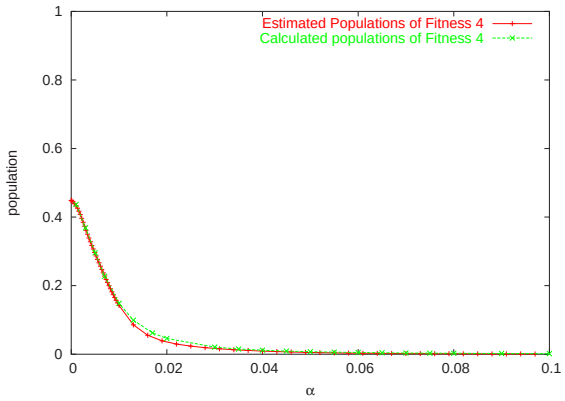
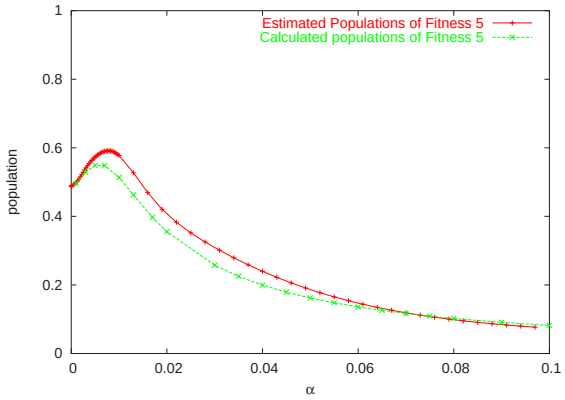


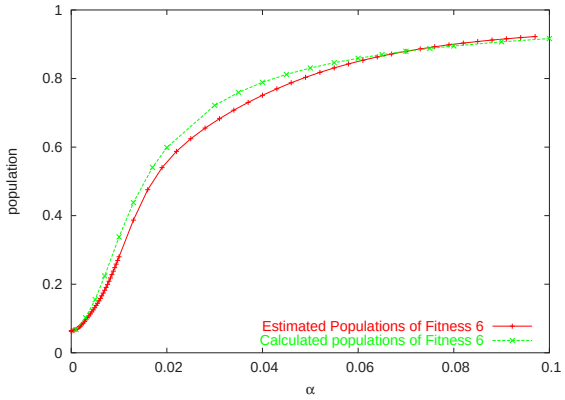
Figure 5.15: 120 member landscape of three dimensional 25-mer functional model proteins, drawn with the same convention as figure 4.3.



(a)



(b)



(c)

Figure 5.16: Steady state populations of (a) fitness 4, (b) fitness 5 and (c) fitness 6 for the 120-member landscape (shown in figure 5.15). Populations are calculated through the population dynamics method and estimated using the free-energy method described in the text.

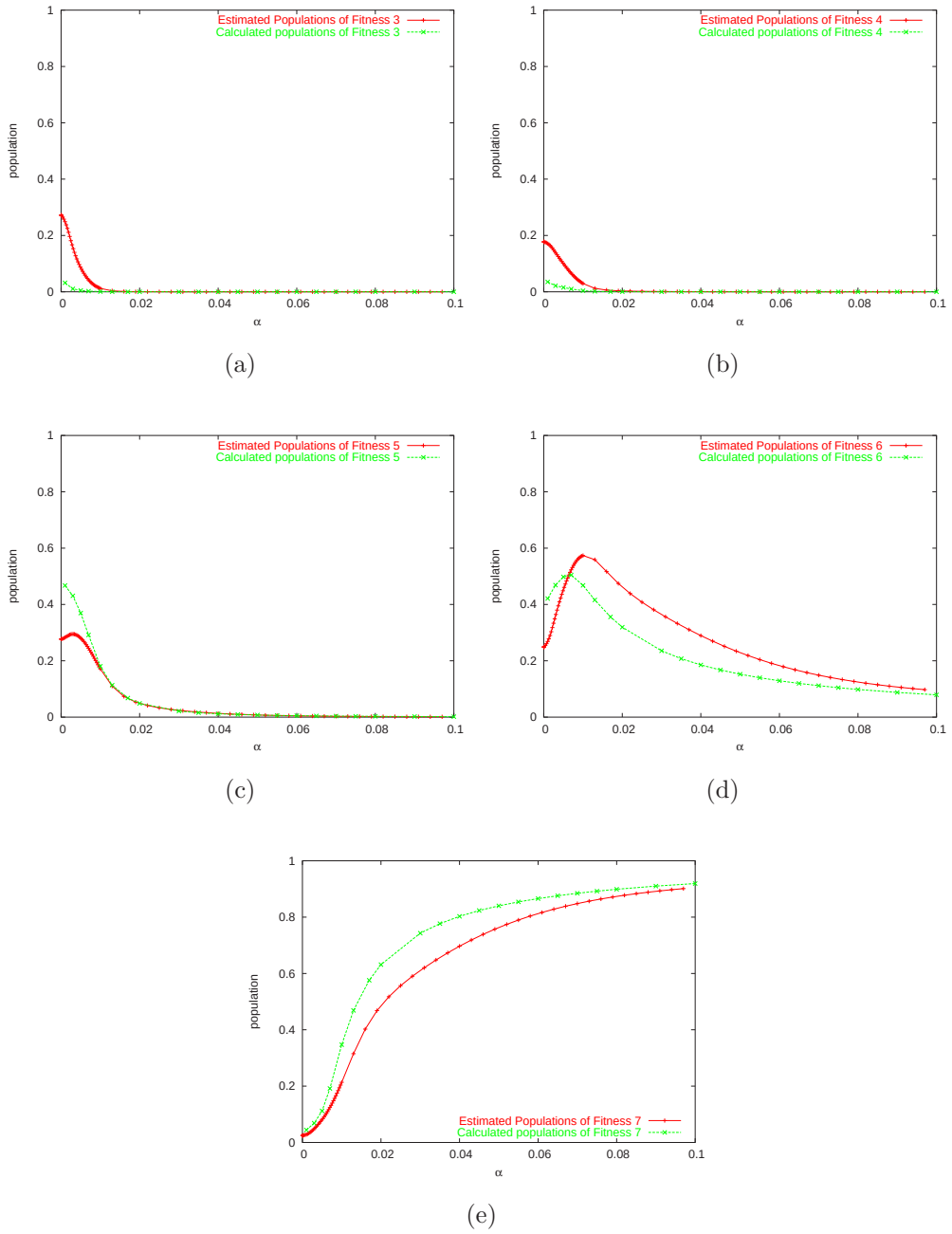


Figure 5.17: Steady state populations of (a) fitness 3, (b) fitness 4, (c) fitness 5, (d) fitness 6 and (e) fitness 7 for the 86-member landscape (shown in figure 5.4). Populations are calculated through the population dynamics method and estimated using the free-energy method described in the text.

connected internally. One sub-landscape is larger than the other. In these respects, this landscape is a toy version of the 86-member landscape. The final populations are given in the label for each node.

The extra node on one side of the landscape is enough to make $> 90\%$ of the population shift to that side. This demonstrates that differences between two sub-landscapes distort the populations away from what would be expected by consideration of their structures. Put another way, alteration of the structure of one sub-landscape will appear to have minimal direct effects on the entropic factors in the other sub-landscape, and will not be captured by the local entropy function. However, because populations on those sub-landscapes are still in direct competition with each other, a strong effect is observed. To illustrate this further consider the removal of the dotted edge in figure 5.18. This will result in two disconnected landscapes. If we maintain them as one landscape for the purposes of a population dynamics simulation, with a starting population present on each landscape, then the population on the smaller landscape will be zero in the steady state. This is consistent with the expectation that the larger landscape will be more mutationally stable.

5.7.2 Rate of convergence under neutral evolution

We observe that, under the population dynamics scheme given earlier, different arbitrary initial populations result in convergence to the same steady state. The number of generations required to reach the steady state serves as a measure of the speed of evolution from the different sequences. The

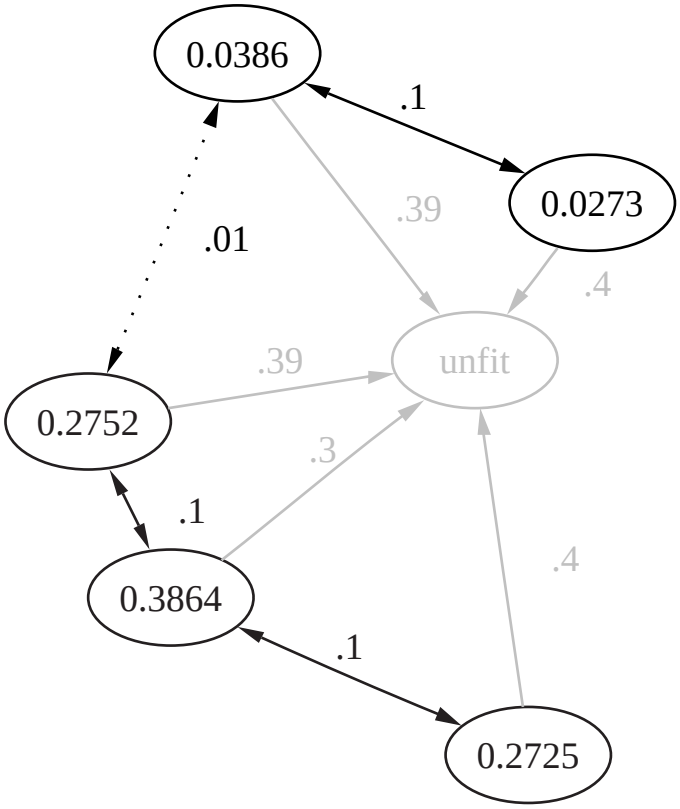


Figure 5.18: Toy landscape to illustrate aspects of population dynamics, described in the body of the text. Edge labels indicate the proportion of population that mutates in that direction at each generation.

steady state is found by running a simulation until $\delta_p < 10^{-20}$. We consider the population dynamics to have converged on subsequent runs when the population of each sequence is within 0.001 of the final population. We calculate population dynamics for $\alpha = 0.3$ and $\mu = 0.01$. These values will result in the population moving to the most fit area of the landscape. We can then consider the time taken to evolve to highest fitness from different areas of the landscape. At the beginning of each simulation a chosen sequence is seeded with a population of one and the time taken to converge to the steady state is recorded. The results (Figure 5.19) show a partitioning of sequences into those which take 5480-5503 generations, 6666-6669 generations and 12504-12507 generations to reach the steady state.

As soon as a sequence of fitness seven is populated it easily out-competes the other sequences and its population rapidly increases. There is then neutral evolution among the three sequences as the hub node becomes the most populous sequence due to the network topology. The time taken to do this depends on which of these nodes is initially the most populous.

In figure 5.20 we consider the change in population for each of these three classes of starting point. A representative sequence from each was chosen, and run for 14 000 generations. The populations of the three "peak" sequences are plotted. In the fastest case, the first node to be reached is the hub node. The majority (64) of the sequences are in this group. The slowest case is where one of the two other sequences acquires a significant population first. A longer period of neutral evolution then occurs as most of the population in the steady state occupies the hub node sequence. The intermediate case is where neighbouring sequences contribute to both the

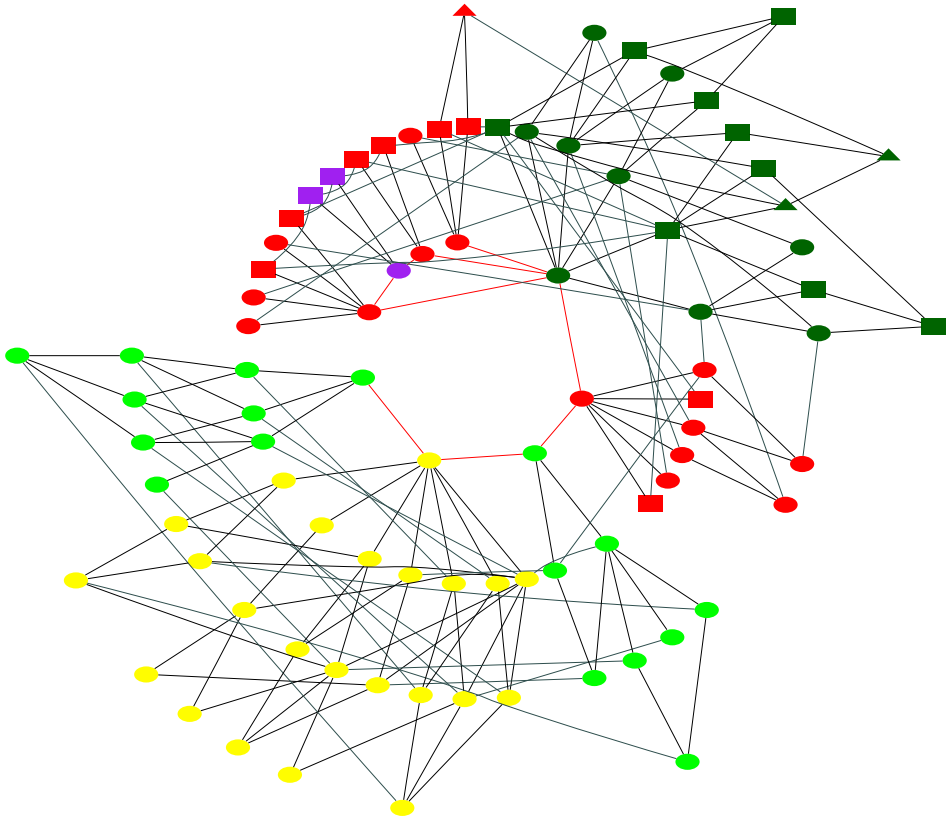
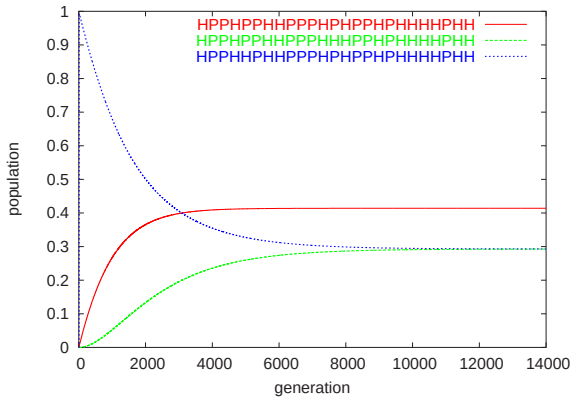
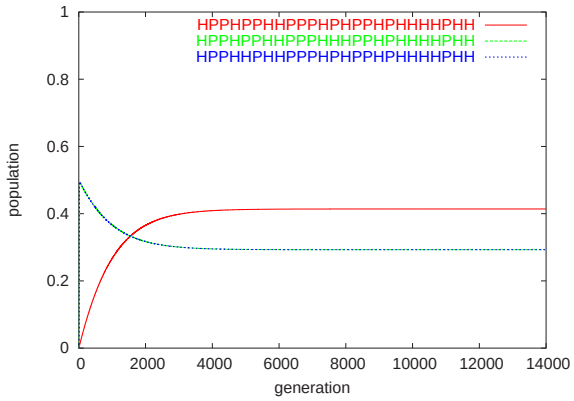


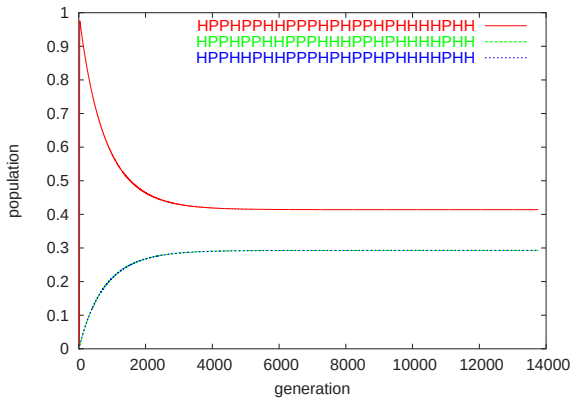
Figure 5.19: 86-member landscape of 25-mers. Node colour indicates fitness. Yellow=3, light green=4, dark green=5, red=6, purple=7. Shape indicates time taken to reach the steady state at $\alpha = 0.3$, $\mu = 0.01$. Ellipse indicates 5480-5503 generations, triangle 6666-6669 generations and rectangle 12504-12507 generations.



(a) Slow Convergence



(b) Medium Convergence



(c) Fast Convergence

Figure 5.20: Population of each sequence of fitness seven over the course of a population dynamics simulation, for starting sequences leading to (a) slow (b) medium and (c) fast convergence. In (b) and (c) the populations of the sequences given in green and blue are identical.

non-hub peak sequences equally. We can visualise the topological reasons for this by excluding the rapidly-evolving nodes from the graph (figure 5.21). Those either side of the triangular nodes contribute the bulk of their populations to the node of highest fitness closest to them (shown in purple). The triangular nodes are equidistant from the nodes of greatest fitness, and so contribute population nearly equally to both nodes of highest fitness. This has the effect of increasing the rate with which the fittest hub node is able to acquire population, and so the system reaches the steady state faster.

For the 151-member landscape the situation is more complex, with a group of 75 sequences taking 7398-7404 generations, 5 groups of 5-15 sequences, and 25 groups of 1-3 sequences. These data were generated with $\alpha = 0.3$ and $\mu = 0.01$. Nevertheless, we still observe that around half (75) of the 151 sequences undergo neutral evolution to converge at the same rate.

Neutral evolution is an important component of molecular evolution [11]. Neutral drift may be important as a necessary step before new function can be acquired [12]. In this section we have shown how a single mutation in an initial starting sequence can make a large difference to its neutral evolution; under the deterministic population dynamics simulation it results in the donation of its population to a different sequence on the peak. Under some landscapes this may result in a shorter or longer period of neutral drift being necessary before higher fitness can be achieved. The use of a stochastic model with a finite population would be appropriate for further investigation of this aspect, as the infinite population in the deterministic model reduces the problem of neutral drift.

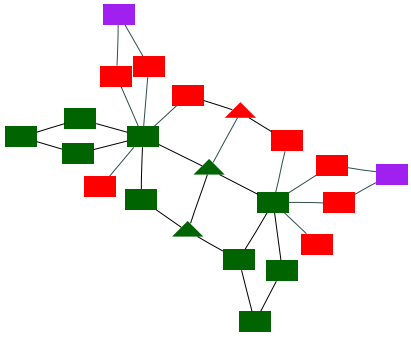


Figure 5.21: 86 member landscape of 25-mers, drawn with the same convention as Figure 5.19, but with the most rapidly evolving nodes excluded.

5.7.3 Rate of convergence under adaptive evolution

Our populations rapidly reach the fittest sequences. Most of the time in previous set of population dynamics simulations involves populations moving between equally fit sequences. This means that the above measure most closely reflects the speed of neutral evolution within a single neutral net. A more pertinent measure is the time taken to reach the peak, excluding the time taken redistributing population on the peak. This reflects the rate of adaptive evolution. This does not mean that no neutral evolution need occur; neutral drift must still take place in order to cross neutral nets and allow adaptive evolution. However, from the view of adaptive evolution, the final neutral drift of sequences of highest possible fitness is unimportant. To this end, we consider the number of generations before 99% of the population occupy the highest available fitness. Figure 5.22 shows this for the 86-member landscape, with $\alpha = 0.3$ and $\mu = 0.01$. The numbers shown in figure 5.22 are smaller by one or two orders of magnitude to those in figure 5.19. On our landscapes, the bulk of adaptive evolution takes place rapidly,

followed by a long period of largely neutral evolution, before reaching a steady state.

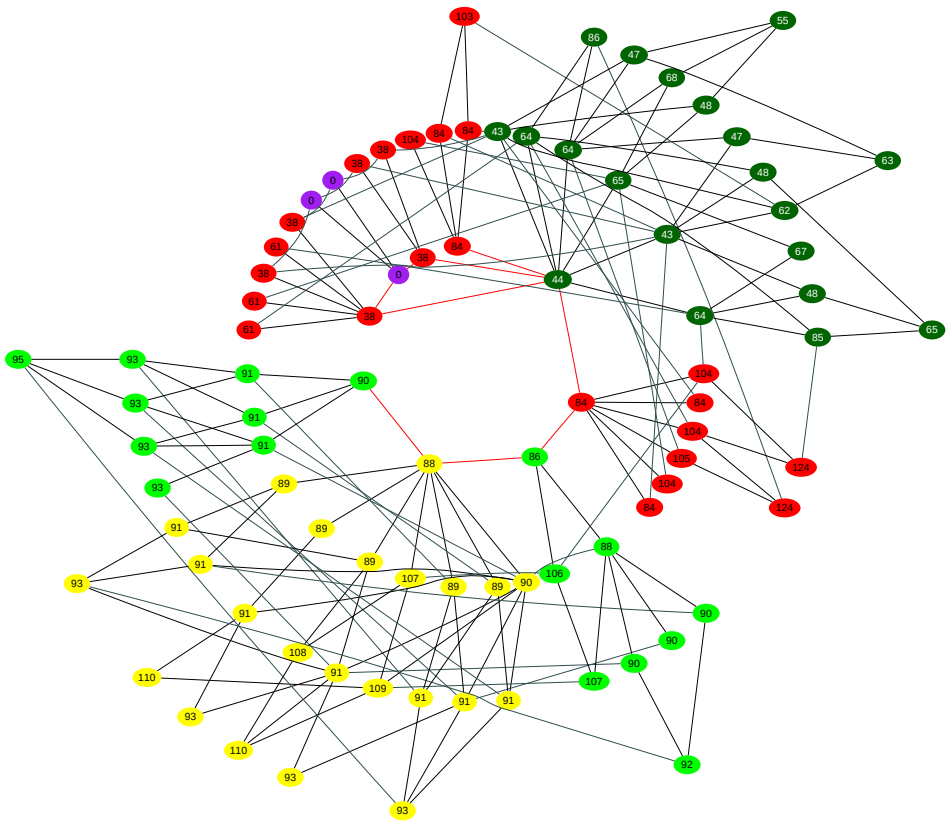


Figure 5.22: 86-member landscape of 25-mers. Colours indicate fitness as in figure 5.4. Numbers indicate the number of generations before 99% of the population occupies the peak (purple nodes).

We observe that there are three "kinetic traps" that can slow evolution in the 86-member landscape. These are neutral nets with fitness six that do not directly link to the landscape peak. One of these must be accessed when populations move through the bottleneck, meaning that the populations of fitness three and four must pass via a kinetic trap, slowing them down. This is borne out by the observation that the number of generations needed to

cross from the areas of fitness three and four is similar to that of fitness six. The region of fitness six acts as a bottleneck, and so the time taken to evolve a significant population of fitness six is negligible in comparison to the time taken to move across to the global maximum of fitness seven.

As demonstrated earlier in figure 4.5, our landscapes display a superfunnel [2] topology. The more thermodynamically stable nodes are those closer to the centre of the graph, and are associated with faster rates of evolution than those further out. We suggest that this is due to the increase in numbers of allowed mutations as stability increases, allowing the more stable sequences greater opportunity to evolve.

Figure 5.23 and Table 5.1 demonstrate how a move towards the hub nodes corresponds to an increase in the rate of adaptive evolution. The vast majority of the edges that change rate show that the rate increases towards the hub nodes. Thus, increasing the thermodynamic stability of a sequence will likely result in more rapid adaptive evolution within a superfunnel.

5.7.4 Effect of hub nodes

Fitness distribution of evolving populations

In order to examine the effect the structure of the landscape has on the population dynamics, we examine the difference in population dynamics when two different starting points are chosen. We examine two runs of our population dynamics algorithm after assigning all the population of the first

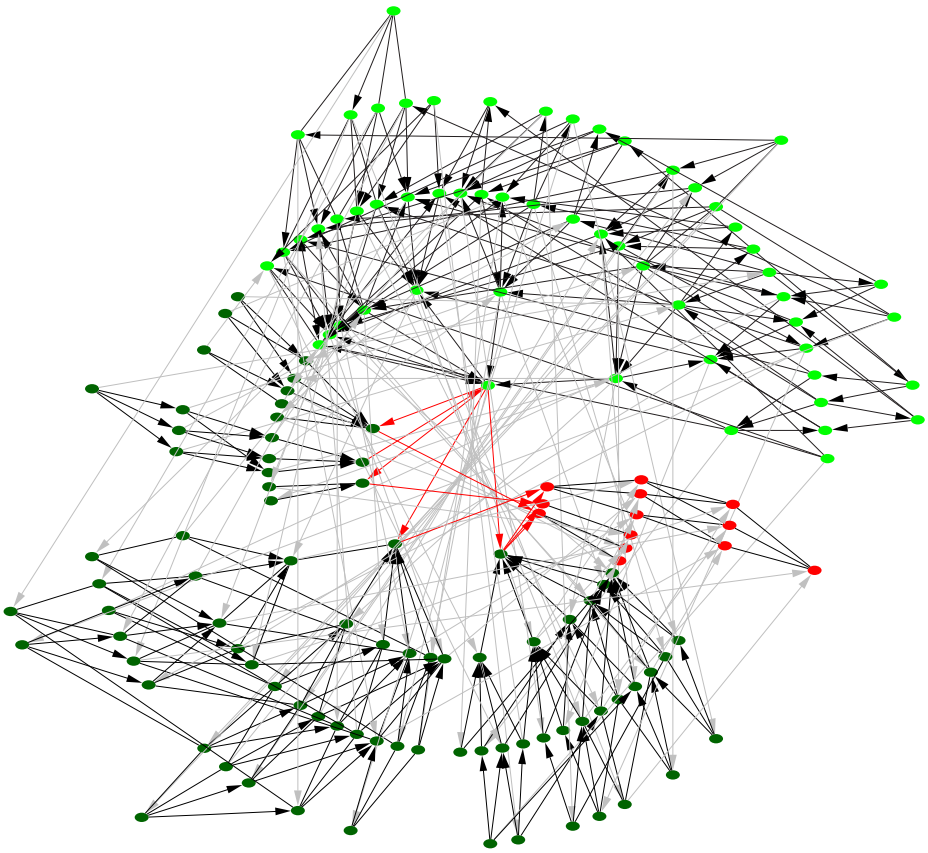


Figure 5.23: 151 member landscape of three dimensional 25-mer functional model proteins. Arrow direction indicates increased rate of adaptive evolution, otherwise the convention is the same as figure 4.3.

Edges that increase rate of adaptive evolution and move closer to hub nodes	189
Edges that decrease rate of adaptive evolution and move closer to hub nodes	24
Edges which maintain distance from hub nodes	107
Edges that maintain rate of adaptive evolution	106
Edges that maintain both rate of adaptive evolution and distance	9

Table 5.1: Effect of movement of the seeding sequence towards the hub nodes on the rate of adaptive evolution.

generation to either a hub node or an outer node with fitness four on the 86-member landscape, with $\mu = 0.01$, $\alpha = 0.3$. Figure 5.24 shows the fraction of population on the peak, over 100 generations of population dynamics simulation for different starting conditions. The effect of starting evolution from a hub node is compared with that of an edge node. The effect is very similar to an increase in μ from 0.01 to 0.017. For larger values μ the hub nodes have a lesser effect (data not shown). This is to be expected, as there is a diminishing return on increasing μ.

5.8 Conclusions

We have considered how the structure of some evolutionary landscapes allows population dynamics to proceed in a manner akin to chemical thermodynamics, allowing estimation of steady state populations from a definition of "free energy". The "entropic" factor in this process is not the number of

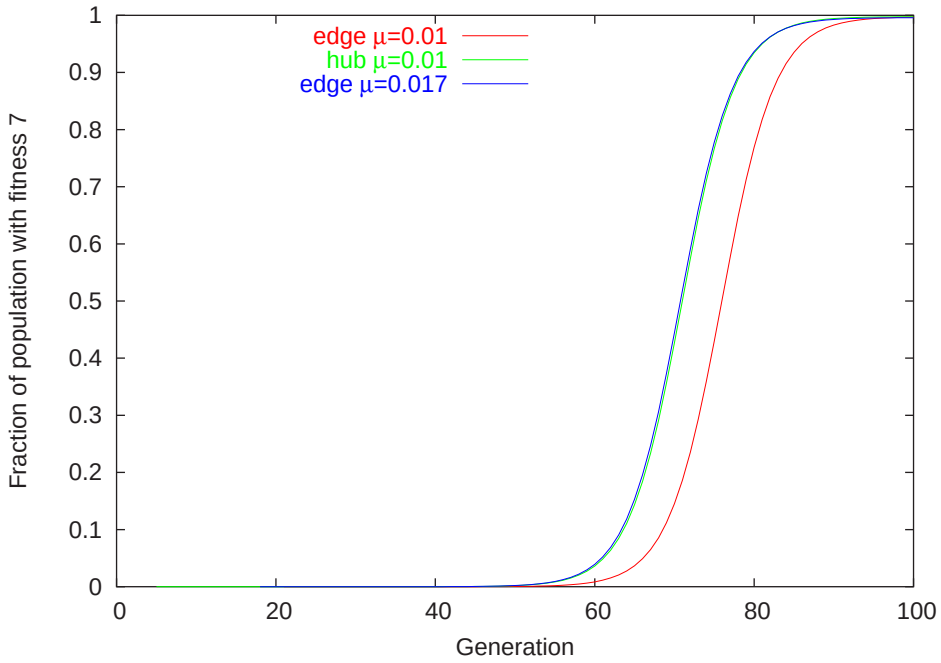


Figure 5.24: Mean population of the fittest sequences over the time course of a 100 population dynamics simulations for different values of μ and seeding sequences.

sequences of a particular fitness, rather it is the connectivity of the landscape.

We also observe landscapes where the connectivity lowers the quantitative accuracy of predictions from this scheme, although the shape of the population changes under changing selection pressures are still qualitatively accurate, predicting rising and falling populations. Why are predictions qualitatively good but quantitatively poor? It seems that the division of a landscape into two weakly linked sub-landscapes is responsible, increasing the relative population of one part of the landscape over what the partition function is able to predict. This may be a consequence of the formulation of the population dynamics process presented in this thesis. This process highly favours parts of the landscape where only a small advantage in mu-

tational robustness is given.

Although we do not use the partition function method to predict populations of individual sequences, it is easy to do this for populations under strong selection. These populations are located almost exclusively at the landscape peak, where they undergo neutral evolution to a steady state distributed over that peak. In these cases, the populations can be estimated accurately by ignoring the sequences that lie off the peak when the next nearest neighbours are counted. The reason these sequences should be neglected is that they account for only a tiny proportion of the donated population onto the peak in the steady state.

For populations under weak selection the partition function works well at predicting individual sequences, as long as the landscape is not split into weakly connected sub-lanscapes. For sequences under intermediate selection, the process is observed to work well for averaged populations, but less well for individual sequences.

The superfunnel structure of the evolutionary landscapes leads to the hub nodes being the most favourable according to our formulation of local entropy. This leads to the over-representation of these sequences during the course of evolution. However, the sheer number of nodes near the rim of the landscape causes the hub node populations to be low. In practice, on larger, more realistic landscapes, we would expect the hub nodes to be rarely populated. This suggests that natural proteins will have thermodynamic and mutational stabilities akin to the rim nodes.

It is possible that directed evolution processes, which attempt to find new

function from proteins by selection and screening of libraries of sequences [7, 15, 13], could benefit from further understanding of the structure of evolutionary landscapes. The superfunnel structure offers one such possibility to improve such strategies; improved stability of a sequence may offer an easier path to improve an existing function, as sequences move closer to the centre of the funnel and so a greater number of acceptable mutations become available to them. One argument against such a strategy might be that the marginal stability of proteins is a required part of their function; however this may not be true [17].

Do natural processes in evolution take advantage of the structure of superfunnels in improving function? From our population dynamics experiments it seems unlikely, as neutral drift is likely to move populations rapidly away from the hub and towards the rim of the superfunnel. Larger landscapes for longer sequences are likely to be more extreme than the landscapes seen here. However, these studies do not not address changes in mutation rate, for example by bacteria under the "SOS response" [5]. It may be that a whole series of neutral mutations is necessary to move the population towards the hub before mutation to new function can take place. This represents a kind of "entropic" effect, akin to the free energy barrier in protein folding where a large number of conformations must be sampled in order to reach the transition state ensemble (see section 1.1.5). A higher mutation rate would be analogous to a higher temperature, allowing sequences to move more rapidly to the hub. A subsequent lower mutation rate on reaching the hub would allow fixation of improved function. Perhaps the study of time-dependent mutation population dynamics [10] on our landscapes could lead to some insight into this possible process.

Bibliography

[1] L.D. Bogarad and M.W. Deem. A hierarchical approach to protein molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.*, 96:2591–2595, 1999.

[2] E. Bornberg-Bauer and H.S. Chan. Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. *Proc. Natl. Acad. Sci. U. S. A.*, 96:10689–10694, 1999.

[3] Y. Cui, W.H. Wong, E. Bornberg-Bauer, and H.S. Chan. Recombinatoric exploration of novel folded structures: A heteropolymer-based model of protein evolutionary landscapes. *Proc. Natl. Acad. Sci. U. S. A.*, 99:809–814, 2002.

[4] K.A. Dill, S. Bromberg, K.Z. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, and H.S. Chan. Principles of protein-folding - a perspective from simple exact models. *Protein Sci.*, 4:561–602, 1995.

[5] H. Echols and M.F. Goodman. Mutation induced by DNA damage - a many protein affair. *Mutat. Res.*, 236:301–311, 1990.

[6] S.I. Grossman and J.E. Turner. *Mathematics for the Biological Sciences*. Macmillan, 1974.

[7] P. Kast and D. Hilvert. 3d structural information as a guide to protein engineering using genetic selection. *Curr. Opin. Struct. Biol.*, 7:470–479, 1997.

[8] S. Kauffmann. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, Oxford, 1993.

[9] T.B. Kepler and A.S. Perelson. Somatic hypermutation in B-cells - an optimal-control treatment. *J. Theor. Biol.*, 164:37–64, 1993.

[10] T.B. Kepler and A.S. Perelson. Modeling and optimization of populations subject to time-dependent mutation. *Proc. Natl. Acad. Sci. U. S. A.*, 92:8219–8223, 1995.

[11] M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, 1983.

[12] D.J. Lipman and W.J. Wilbur. Modeling neutral and selective evolution of protein folding. *Proc. R. Soc. Lond. Ser. B-Biol. Sci.*, 245:7–11, 1991.

[13] J. Minshull and W.P.C. Stemmer. Protein evolution by molecular breeding. *Curr. Opin. Chem. Biol.*, 3:284–290, 1999.

[14] A.S. Perelson and C.A. Macken. Protein evolution on partially correlated landscapes. *Proc. Natl. Acad. Sci. U. S. A.*, 92:9657–9661, 1995.

[15] Z.X. Shao and F.H. Arnold. Engineering new functions and altering existing functions. *Curr. Opin. Struct. Biol.*, 6:513–518, 1996.

[16] D.M. Taverna and R.A. Goldstein. The distribution of structures in evolving protein populations. *Biopolymers*, 53:1–8, 2000.

[17] D.M. Taverna and R.A. Goldstein. Why are proteins marginally stable? *Proteins*, 46:105–109, 2002.

[18] Y. Xia and M. Levitt. Roles of mutation and recombination in the evolution of protein thermodynamics. *Proc. Natl. Acad. Sci. U. S. A.*, 99:10382–10387, 2002.

[19] T. Yomo, S. Saito, and M. Sasai. Gradual development of protein-like global structures through functional selection. *Nat. Struct. Biol.*, 6:743–746, 1999.

Chapter 6

Conclusions

This thesis has examined the nature of protein evolutionary landscapes, from the perspective of simple lattice model proteins. The model has been extended to longer chains than have previously been examined, with exhaustive enumeration, through optimisations of the searches of sequence and structure space. The model was then extended to the three-dimensional diamond lattice. This adds some realism, as obviously proteins are three-dimensional, at minimal additional cost; both the square and the diamond lattices are four-coordinate.

The two dimensional model proteins were examined exhaustively up to length $n = 23$. Their landscapes become larger and more complex with increasing chain length, with larger fractions of neutral mutations. The importance of neutral mutations for adaptive evolution has been considered in earlier work [4]. However, that model regarded adaptive mutations as those which alter structure. In our model, mutations may keep the same

structure, and yet alter the fitness. This is a more realistic representation of real protein sequences, and neutrality is still shown to be an important component of evolution in our model. Stretches of neutral landscape must be crossed in order to gain new function. We observe that this is a more important aspect of longer sequences, and must be expected to have increased importance for larger alphabets. The implication of this is that, for more protein-like models, long periods of evolution are spent in neutral drift. This represents one form of an "entropic" effect that opposes adaptive evolution. The forces of selection are unable to drive a population towards a more fit area of the landscape as time must be spent sampling many different sequences. This is analogous to protein folding, where the initial barrier is entropic, with large numbers of conformations being sampled until an appropriate intermediate is found.

In the case of two-state protein folding, once the sequence is folded into a transition state its folding is fast. However, we observe features of the evolutionary landscape that may introduce an additional opposing force. In the square lattice models, we observe critical edges, parts of the landscape connected by a single path of mutations. All mutations off the path are lethal. In the diamond lattice model, this feature is not so pronounced. However, similar bottleneck features are observed, for example on the 86-member landscape of 25-mers. A population may drift towards an area of higher fitness; however, there may be a bottleneck to pass before higher fitness can be achieved. A population on a bottleneck will be easily outcompeted by a population on a flat landscape of equal fitness. This is the quasispecies theory of Eigen [3]; due to the constant production of mutant sequences, selection acts not on single sequences but on a "cloud" of related

sequences. A population crossing a bottleneck will be disfavoured as each sequences is closely related to many more non-fit sequences than on a flat landscape. However, once the bottleneck is crossed there may be a region of higher fitness, and the population will have a competitive advantage.

A complementary feature of landscapes is considered for the diamond-lattice model. Superfunnels [2] are observed - the most stable sequences are rare, and landscapes are structured such that increasing mutational distance from the most stable sequence results in decreasing stability. Each neutral net has a set of neighbouring neutral nets - those connected to one or more of its sequences by a single point mutation. The most stable sequence in each neutral net is found to be a hub node; it is able to mutate to all the neighbouring neutral nets. This is a consequence of the increased number of interconnections available as we descend the superfunnel; more stable sequences are able to sustain more mutations in the binding pocket. Our landscapes predict that increased stability leads to increased potential for favourable adaptive mutation. This may have implications for directed evolution. The converse principle, that decreased stability results in decreased evolutionary potential, could be useful in consideration of drug resistance and cancer. In these cases, it would be useful to prevent mutation to increased fitness, for example, to prevent increased activity of an oncogene in the production of cancer. It may be that critical edges exist most frequently where the landscape is least stable. Crossing a critical edge or bottleneck may result in a change in protein structure. Hence, the sequences around the critical edges would be expected to be very marginally stable. In this way, sequence space may consist of a set of superfunnels corresponding to single structures, linked together by evolutionary bottlenecks. There are al-

ready results that suggest this structure [1], and this would be an interesting area for further investigation with longer chains.

If this structure of sequence space is an accurate representation, then it may be necessary to extend the model to study it. Currently, fitness is defined independent of structure. It may be that optimisation within a superfunnel, best achieved by the most stable sequences, reflects a refinement of function, for example, an increase in rate of catalysis by an enzyme. However, attainment of a new structure is best achieved by less stable sequences. This may correspond to attainment of a new function, e.g. enzyme catalysis of a new reaction.

We conclude our work with a study of population dynamics on the landscapes. The simulations show how populations tend to drift away from the hub nodes and towards the rim of the landscape. Furthermore, populations can be predicted from a simple scheme based on the partition function, when neutral nets are characterised by their fitnesses and a measure of their connection to the landscape, the number of next-nearest neighbours. That this calculation is successful highlights the importance of landscape structure on population.

Bibliography

[1] E. Bornberg-Bauer. Randomness, structural uniqueness, modularity and neutral evolution in sequence space of model proteins. *Z. Phys. Chemie-Int. J. Res. Phys. Chem. Chem. Phys.*, 216:139–154, 2002.

[2] E. Bornberg-Bauer and H.S. Chan. Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. *Proc. Natl. Acad. Sci. U. S. A.*, 96:10689–10694, 1999.

[3] M. Eigen. Selforganization of matter and the evolution of macromolecules. *Naturwissenschaften*, 10:465–523, 1971.

[4] D.J. Lipman and W.J. Wilbur. Modeling neutral and selective evolution of protein folding. *Proc. R. Soc. Lond. Ser. B-Biol. Sci.*, 245:7–11, 1991.

Chapter 7

Appendix

We include the sequence data for the three largest landscapes for the 25-mer on the diamond lattice. These sequences are referred to in much of the text, and may be useful for those who seek to reproduce these results. Hydrophobes are given as 1, polar residues as 0. The population of the native state is given from the method described in section 4.5.

7.1 86 member landscape

sequence	fitness	native state energy	1st excited state energy	population of native state
100100010000010101000011	3	-14	-12	0.9989
100100010000010101000111	4	-14	-12	0.9985
100100010000010101001011	4	-14	-12	0.9985
100100010000010101010011	3	-14	-12	0.9988
100100010000011101000011	3	-14	-12	0.9978
100100010000011101000111	4	-14	-12	0.9967
100100010000011101001011	4	-14	-12	0.9966
100100010000011101010011	3	-14	-12	0.9965
100100110000010101000011	3	-14	-12	0.9979
100100110000010101000111	4	-14	-12	0.9971
100100110000010101001011	4	-14	-12	0.9955
100100110000010101010011	3	-14	-12	0.9968
100100110000011101000011	3	-14	-12	0.9937
100100110001010010101001011	5	-16	-14	0.9988
100100110001010010101011011	6	-16	-14	0.9970
100100110001010010101101011	6	-16	-14	0.9975
100100110001010010101111011	7	-16	-14	0.9914
100100110001010010111001011	5	-16	-14	0.9963
100100110001010010111011011	6	-16	-14	0.9914
100100110001010110101001011	6	-16	-14	0.9981
100100110001011010101001011	6	-16	-14	0.9963
100100110001011010111001011	6	-16	-14	0.9888
100100110001111001010011	5	-16	-14	0.9974
100100110001111001010111011	6	-16	-14	0.9947

10010011000111001011010 11	6	-16	-14	0.9892
10010011000111001011110 11	7	-16	-14	0.9744
10010011000111011010010 11	6	-16	-14	0.9915
10010011000111101010010 11	6	-16	-14	0.9913
10010101000001101010000 11	3	-14	-12	0.9981
10010101000001101010010 11	4	-14	-12	0.9953
10010101000001101010100 11	3	-14	-12	0.9970
10010101000001111010000 11	3	-14	-12	0.9946
10010111000001101010000 11	3	-14	-12	0.9940
10010111000001101010010 11	4	-14	-12	0.9857
10010111000001111010000 11	3	-14	-12	0.9838
10010111000101001010010 11	5	-16	-14	0.9972
10010111000101001010110 11	6	-16	-14	0.9924
10010111000101001110010 11	5	-16	-14	0.9912
10010111000101101010010 11	6	-16	-14	0.9915
10010111000101101110010 11	6	-16	-14	0.9761
10011001000001101010000 11	3	-14	-12	0.9980
10011001000001101010001 11	4	-14	-12	0.9972
10011001000001101010010 11	4	-14	-12	0.9961
10011001000001111010000 11	3	-14	-12	0.9939
10011011000001101010000 11	3	-14	-12	0.9969
10011011000001101010001 11	4	-14	-12	0.9956
10011011000101001010010 11	5	-16	-14	0.9963
10011011000101001010110 11	6	-16	-14	0.9918
10011011000101001011010 11	6	-16	-14	0.9908
10011011000101001011110 11	7	-16	-14	0.9789
10011011000101001110010 11	5	-16	-14	0.9906
10011011000101011010010 11	6	-16	-14	0.9925

10011011000101101010010 11	6	-16	-14	0.9921
10011011000111001010010 11	5	-16	-14	0.9905
10011011000111011010010 11	6	-16	-14	0.9723
10011111000101001010010 11	5	-16	-14	0.9896
10011111000101001110010 11	5	-16	-14	0.9779
10110001000001101010000 11	3	-14	-12	0.9982
10110001000001101010001 11	4	-14	-12	0.9975
10110011000001101010000 11	3	-14	-12	0.9964
10110011000001101010001 11	4	-14	-12	0.9947
10110011000101001010010 11	5	-16	-14	0.9969
10110011000101001010110 11	6	-16	-14	0.9927
10110011000101011010010 11	6	-16	-14	0.9917
10110011000101101010010 11	6	-16	-14	0.9924
10110011000111001010010 11	5	-16	-14	0.9936
10110111000101001010010 11	5	-16	-14	0.9916
10111001000001101010000 11	3	-14	-12	0.9969
10111001000001101010001 11	4	-14	-12	0.9958
10111011000001101010000 11	3	-14	-12	0.9947
10111011000001101010001 11	4	-14	-12	0.9926
11010001000001101010000 11	3	-14	-12	0.9984
11010001000001101010010 11	4	-14	-12	0.9972
11010001000001111010000 11	3	-14	-12	0.9972
11010001000001111010010 11	4	-14	-12	0.9942
11010011000001101010000 11	3	-14	-12	0.9964
11010011000001111010000 11	3	-14	-12	0.9890
11010011000101001010010 11	5	-16	-14	0.9983
11010011000101001110010 11	5	-16	-14	0.9926
11010011000101101010010 11	6	-16	-14	0.9938

11010011000101101111001011	6	-16	-14	0.9793
11010011000111001010011011	5	-16	-14	0.9952
11011011000101001010011011	5	-16	-14	0.9932
11011011000111001010011011	5	-16	-14	0.9791
11110011000101001010011011	5	-16	-14	0.9939
11110011000111001010011011	5	-16	-14	0.9848

7.2 120 member landscape

sequence	fitness	native state energy	1st excited state energy	population of native state
100010001001000001101010	4	-14	-12	0.9998
100010001001000001101110	4	-14	-12	0.9998
100010001001000001111010	4	-14	-12	0.9997
100010001001000001111110	4	-14	-12	0.9997
100010001001000101101010	5	-14	-12	0.9991
100010001001000101101110	5	-14	-12	0.9975
100010001001000101111010	5	-14	-12	0.9980
100010001001000101111110	5	-14	-12	0.9964
100010001001001011101010	5	-14	-12	0.9989
100010001001001011101110	5	-14	-12	0.9973
100010001001001011111010	5	-14	-12	0.9979
100010001001001011111110	5	-14	-12	0.9959
100010001001010011101010	5	-14	-12	0.9993
100010001001010011101110	5	-14	-12	0.9978
100010001011000011101010	4	-14	-12	0.9997
100010001011000011101110	4	-14	-12	0.9997
100010001011000011111010	4	-14	-12	0.9996
100010001011000011111110	4	-14	-12	0.9996
100010001011001011101010	5	-14	-12	0.9987
100010001011001011101110	5	-14	-12	0.9961
100010001011001011111010	5	-14	-12	0.9977
100010001011001011111110	5	-14	-12	0.9946
100010001101000011101010	4	-14	-12	0.9983
100010001101000011101110	4	-14	-12	0.9931

10001000111100000011101010	4	-14	-12	0.9948
10001000111100000011101110	4	-14	-12	0.9840
10001001100100000011101010	4	-14	-12	0.9993
10001001100100000011101110	4	-14	-12	0.9982
10001001100100000011111010	4	-14	-12	0.9988
10001001100100001011101010	5	-14	-12	0.9895
10001001110100000011101010	4	-14	-12	0.9977
10001001110100000011101110	4	-14	-12	0.9901
10001010100100000011101010	4	-14	-12	0.9993
10001010100100000011101110	4	-14	-12	0.9988
10001010100100000011111010	4	-14	-12	0.9993
10001010100100000011111110	4	-14	-12	0.9987
10001010100100001011101010	5	-14	-12	0.9954
10001010101100000011101010	4	-14	-12	0.9986
10001010101100000011101110	4	-14	-12	0.9978
10001010101100000011111010	4	-14	-12	0.9986
10001010101100000011111110	4	-14	-12	0.9977
10001011100100000011101010	4	-14	-12	0.9984
10001011100100000011101110	4	-14	-12	0.9951
10001100100100000011101010	4	-14	-12	0.9985
10001100100100000011101110	4	-14	-12	0.9965
10001100100100000011111010	4	-14	-12	0.9968
10001100100100000011111110	4	-14	-12	0.9934
10001100100100001011101010	5	-14	-12	0.9904
10001100101100000011101010	4	-14	-12	0.9964
10001100101100000011101110	4	-14	-12	0.9923
10001100101100000011111010	4	-14	-12	0.9940
10001100101100000011111110	4	-14	-12	0.9882

10001100110100000011101010	4	-14	-12	0.9951
10001100110100000011101110	4	-14	-12	0.9826
10001100111100000011101010	4	-14	-12	0.9842
10001100111100000011101110	4	-14	-12	0.9582
10001101100100000011101010	4	-14	-12	0.9958
10001101110100000011101010	4	-14	-12	0.9917
10001110100100000011101010	4	-14	-12	0.9945
10001110100100000011101110	4	-14	-12	0.9891
10001110100100000011111010	4	-14	-12	0.9916
10001110100100000011111110	4	-14	-12	0.9844
10001111100100000011101010	4	-14	-12	0.9893
10011000100100000011101010	5	-14	-12	0.9993
10011000100100000011101110	5	-14	-12	0.9987
10011000100100000011111010	5	-14	-12	0.9978
10011000101100000011101010	5	-14	-12	0.9979
10011000101100000011101110	5	-14	-12	0.9963
10011000110100000011101010	5	-14	-12	0.9961
10011000110100000011101110	5	-14	-12	0.9878
10011000111100000011101010	5	-14	-12	0.9873
10011000111100000011101110	5	-14	-12	0.9662
10011001100100000011101010	5	-14	-12	0.9972
10011001100100000011101110	5	-14	-12	0.9935
10011001100100000011111010	5	-14	-12	0.9931
10011001110100000011101010	5	-14	-12	0.9926
10011001110100000011101110	5	-14	-12	0.9789
10011100100100000011101010	5	-14	-12	0.9964
10011100100100000011101110	5	-14	-12	0.9905
10011100100100000011111010	5	-14	-12	0.9915

10011100101100000011101010	5	-14	-12	0.9913
10011100101100000011101110	5	-14	-12	0.9797
10011100110100000011101010	5	-14	-12	0.9906
10011100110100000011101110	5	-14	-12	0.9714
10011100111100000011101010	5	-14	-12	0.9712
10011100111100000011101110	5	-14	-12	0.9272
10011101100100000011101010	5	-14	-12	0.9919
10011101110100000011101010	5	-14	-12	0.9843
11001000100100000011101010	5	-14	-12	0.9993
11001000100100000011101110	5	-14	-12	0.9982
11001000100100000011111010	5	-14	-12	0.9978
11001000110100000011101010	5	-14	-12	0.9965
11001000110100000011101110	5	-14	-12	0.9833
11001001100100000011101010	5	-14	-12	0.9987
11001001100100000011101110	5	-14	-12	0.9955
11001001100100000011111010	5	-14	-12	0.9958
11001001110100000011101010	5	-14	-12	0.9954
11001001110100000011101110	5	-14	-12	0.9768
11001010100100000011101010	5	-14	-12	0.9983
11001010100100000011101110	5	-14	-12	0.9957
11001010100100000011111010	5	-14	-12	0.9950
11001011100100000011101010	5	-14	-12	0.9966
11001011100100000011101110	5	-14	-12	0.9883
11001100100100000011101010	5	-14	-12	0.9965
11001100100100000011101110	5	-14	-12	0.9893
11001100100100000011111010	5	-14	-12	0.9877
11001100110100000011101010	5	-14	-12	0.9910
11001100110100000011101110	5	-14	-12	0.9664

11001101100100000011101010	5	-14	-12	0.9934
11001101110100000011101010	5	-14	-12	0.9871
11001110100100000011101010	5	-14	-12	0.9873
11001111100100000011101010	5	-14	-12	0.9809
11011000100100000011101010	6	-14	-12	0.9975
11011000110100000011101010	6	-14	-12	0.9928
11011001100100000011101010	6	-14	-12	0.9952
11011001110100000011101010	6	-14	-12	0.9889
11011100100100000011101010	6	-14	-12	0.9932
11011100110100000011101010	6	-14	-12	0.9851
11011101100100000011101010	6	-14	-12	0.9868
11011101110100000011101010	6	-14	-12	0.9769

7.3 151 member landscape

sequence	fitness	native state energy	1st excited state energy	population of native state
010001000100100000111010 1	4	-14	-12	0.9996
010001000100100000111011 1	4	-14	-12	0.9996
010001000100100000111110 1	4	-14	-12	0.9996
010001000100100000111111 1	4	-14	-12	0.9996
010001000100100010111010 1	5	-14	-12	0.9991
010001000100100010111011 1	5	-14	-12	0.9978
010001000100100010111110 1	5	-14	-12	0.9972
010001000100100010111111 1	5	-14	-12	0.9959
010001000100100100111010 1	5	-14	-12	0.9985
010001000100100100111011 1	5	-14	-12	0.9961
010001000100100100111110 1	5	-14	-12	0.9963
010001000100100100111111 1	5	-14	-12	0.9934
010001000100101000111010 1	5	-14	-12	0.9988
010001000100101000111011 1	5	-14	-12	0.9971
010001000101100000111010 1	4	-14	-12	0.9996
010001000101100000111011 1	4	-14	-12	0.9996
010001000101100000111110 1	4	-14	-12	0.9993
010001000101100000111111 1	4	-14	-12	0.9993
010001000101100100111010 1	5	-14	-12	0.9971
010001000101100100111011 1	5	-14	-12	0.9931
010001000101100100111110 1	5	-14	-12	0.9946
010001000101100100111111 1	5	-14	-12	0.9901
010001000110100000111010 1	4	-14	-12	0.9989
010001000110100000111011 1	4	-14	-12	0.9957

01000100011010000001111101	4	-14	-12	0.9986
01000100011010000001111111	4	-14	-12	0.9954
01000100011110000001110101	4	-14	-12	0.9924
01000100011110000001110111	4	-14	-12	0.9857
01000100110010000001110101	4	-14	-12	0.9995
01000100110010000001110111	4	-14	-12	0.9995
01000100110010000001111101	4	-14	-12	0.9990
01000100110010000001111111	4	-14	-12	0.9990
01000100110010000101110101	5	-14	-12	0.9972
01000100110010000101111101	5	-14	-12	0.9938
01000100110010100001110101	5	-14	-12	0.9949
01000100110110000001110101	4	-14	-12	0.9990
01000100110110000001110111	4	-14	-12	0.9981
01000100111010000001110101	4	-14	-12	0.9988
01000100111010000001110111	4	-14	-12	0.9954
01000100111010000001111101	4	-14	-12	0.9977
01000100111110000001110101	4	-14	-12	0.9914
01000100111110000001110111	4	-14	-12	0.9836
01000101010010000001110101	4	-14	-12	0.9992
01000101010010000001110111	4	-14	-12	0.9985
01000101010010000001111101	4	-14	-12	0.9992
01000101010010000001111111	4	-14	-12	0.9985
01000101010010000101110101	5	-14	-12	0.9965
01000101010110000001110101	4	-14	-12	0.9980
01000101010110000001110111	4	-14	-12	0.9968
01000101010110000001111101	4	-14	-12	0.9976
01000101010110000001111111	4	-14	-12	0.9964
01000101110010000001110101	4	-14	-12	0.9985

010001011100100000 1110111	4	-14	-12	0.9966
010001011100100010 1110101	5	-14	-12	0.9902
010001100100100000 1110101	4	-14	-12	0.9989
010001100100100000 1110111	4	-14	-12	0.9977
010001100100100000 1111101	4	-14	-12	0.9977
010001100100100000 1111111	4	-14	-12	0.9945
010001100100100010 1110101	5	-14	-12	0.9954
010001100101100000 1110101	4	-14	-12	0.9973
010001100101100000 1110111	4	-14	-12	0.9950
010001100101100000 1111101	4	-14	-12	0.9951
010001100101100000 1111111	4	-14	-12	0.9907
010001100110100000 1110101	4	-14	-12	0.9975
010001100110100000 1110111	4	-14	-12	0.9906
010001100110100000 1111101	4	-14	-12	0.9952
010001100111100000 1110101	4	-14	-12	0.9844
010001100111100000 1110111	4	-14	-12	0.9701
010001101100100000 1110101	4	-14	-12	0.9979
010001101101100000 1110101	4	-14	-12	0.9936
010001101110100000 1110101	4	-14	-12	0.9963
010001101111100000 1110101	4	-14	-12	0.9791
010001110100100000 1110101	4	-14	-12	0.9963
010001110100100000 1110111	4	-14	-12	0.9926
010001110100100000 1111101	4	-14	-12	0.9939
010001110100100000 1111111	4	-14	-12	0.9880
010001111100100000 1110101	4	-14	-12	0.9932
010011000100100000 1110101	5	-14	-12	0.9995
010011000100100000 1110111	5	-14	-12	0.9990
010011000100100000 1111101	5	-14	-12	0.9975

010011000101100000111010 1	5	-14	-12	0.9986
010011000101100000111011 1	5	-14	-12	0.9964
010011000110100000111010 1	5	-14	-12	0.9974
010011000110100000111011 1	5	-14	-12	0.9918
010011000110100000111110 1	5	-14	-12	0.9929
010011000111100000111010 1	5	-14	-12	0.9855
010011000111100000111011 1	5	-14	-12	0.9696
010011001100100000111010 1	5	-14	-12	0.9976
010011001100100000111011 1	5	-14	-12	0.9961
010011001100100000111110 1	5	-14	-12	0.9929
010011001101100000111010 1	5	-14	-12	0.9927
010011001101100000111011 1	5	-14	-12	0.9865
010011001110100000111010 1	5	-14	-12	0.9946
010011001110100000111011 1	5	-14	-12	0.9872
010011001110100000111110 1	5	-14	-12	0.9845
010011001111100000111010 1	5	-14	-12	0.9775
010011001111100000111011 1	5	-14	-12	0.9541
010011100100100000111010 1	5	-14	-12	0.9975
010011100100100000111011 1	5	-14	-12	0.9937
010011100100100000111110 1	5	-14	-12	0.9921
010011100101100000111010 1	5	-14	-12	0.9926
010011100101100000111011 1	5	-14	-12	0.9824
010011100110100000111010 1	5	-14	-12	0.9940
010011100110100000111011 1	5	-14	-12	0.9823
010011100110100000111110 1	5	-14	-12	0.9827
010011100111100000111010 1	5	-14	-12	0.9722
010011100111100000111011 1	5	-14	-12	0.9397
010011101100100000111010 1	5	-14	-12	0.9945

010011101101100000111010 1	5	-14	-12	0.9834
010011101110100000111010 1	5	-14	-12	0.9901
010011101111100000111010 1	5	-14	-12	0.9599
011001000100100000111010 1	5	-14	-12	0.9995
011001000100100000111011 1	5	-14	-12	0.9986
011001000100100000111110 1	5	-14	-12	0.9978
011001000100100010111010 1	6	-14	-12	0.9982
011001000100101000111010 1	6	-14	-12	0.9973
011001000110100000111010 1	5	-14	-12	0.9978
011001000110100000111011 1	5	-14	-12	0.9895
011001000110100000111110 1	5	-14	-12	0.9944
011001001100100000111010 1	5	-14	-12	0.9992
011001001100100000111011 1	5	-14	-12	0.9981
011001001100100000111110 1	5	-14	-12	0.9960
011001001100100010111010 1	6	-14	-12	0.9961
011001001100101000111010 1	6	-14	-12	0.9932
011001001110100000111010 1	5	-14	-12	0.9973
011001001110100000111011 1	5	-14	-12	0.9877
011001001110100000111110 1	5	-14	-12	0.9916
011001010100100000111010 1	5	-14	-12	0.9987
011001010100100000111011 1	5	-14	-12	0.9965
011001010100100000111110 1	5	-14	-12	0.9954
011001011100100000111010 1	5	-14	-12	0.9972
011001011100100000111011 1	5	-14	-12	0.9921
011001100100100000111010 1	5	-14	-12	0.9984
011001100100100000111011 1	5	-14	-12	0.9952
011001100100100000111110 1	5	-14	-12	0.9910
011001100100100010111010 1	6	-14	-12	0.9923

011001100110100000111010 1	5	-14	-12	0.9960
011001100110100000111011 1	5	-14	-12	0.9829
011001100110100000111110 1	5	-14	-12	0.9850
011001101100100000111010 1	5	-14	-12	0.9972
011001101110100000111010 1	5	-14	-12	0.9944
011001110100100000111010 1	5	-14	-12	0.9924
011001111100100000111010 1	5	-14	-12	0.9881
011011000100100000111010 1	6	-14	-12	0.9988
011011000110100000111010 1	6	-14	-12	0.9956
011011001100100000111010 1	6	-14	-12	0.9968
011011001110100000111010 1	6	-14	-12	0.9925
011011100100100000111010 1	6	-14	-12	0.9964
011011100110100000111010 1	6	-14	-12	0.9919
011011101100100000111010 1	6	-14	-12	0.9929
011011101110100000111010 1	6	-14	-12	0.9872