Population dynamics simulations of functional model proteins

Benjamin P. Blackburne
Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA

Jonathan D. Hirst[a)]
School of Chemistry, University of Nottingham, University Park, Nottingham NG7 2RD

(Received 22 June 2005; accepted 16 August 2005; published online 20 October 2005)

In order to probe the fundamental principles that govern protein evolution, 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 an explicit definition 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 study the fitness landscapes of these functional model proteins, that is, the set of all sequences mapped on to their corresponding fitnesses and connected to their one mutant neighbors. Through population dynamics simulations we directly study the influence of the nature of the fitness landscape on evolution. Populations are observed to move to a steady state, the distribution of which can often be predicted prior to the population dynamics simulations from the nature of the fitness landscape and a quantity analogous to a partition function. In this paper, we develop a scheme for predicting the steady-state population on a fitness landscape, based on the nature of the fitness landscape, thereby obviating the need for explicit population dynamics simulations and providing some insight into the impact on molecular evolution of the nature of fitness landscapes. Poor predictions are indicative of fitness landscapes that consist of a series of weakly connected sublandscapes. © *2005 American Institute of Physics*. [DOI: 10.1063/1.2056545]

I. INTRODUCTION

In order to gain insights into molecular evolution, various models have been proposed.[1] One of the most popular is the lattice model of proteins[2] initially applied to the study of protein folding.[3] Lattice models, also known as minimalist models, can permit complete sampling of all sequences or structures at the expense of representation of fine molecular details. For the examination of the generic aspects of protein evolution, it may be advantageous to consider minimalist models because unnecessary atomistic detail may obscure the general principles that might emerge from the simulations.[4] More recently, lattice models have been applied to the characterization of the protein fitness landscapes[5,6] and the study of the effect of the structure of these networks on evolving populations.[7,8]

In lattice models, individual amino acids are represented as beads on a chain, which is restricted to lie on a lattice. Typically a square lattice is used,[3,5] although other lattices such as cubic[9] or diamond[6] may be employed. In the case of the diamond lattice, the chain is three dimensional and the bond angles between beads are 109°. 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, that is, beads that are nearest neighbors on the lattice in that particular conformation, but not nearest neighbors in the chain.

The fitness landscape so generated can be characterized by the use of population dynamics simulations. The landscape is a description of sequence space, the allowed transitions between sequences (i.e., which sequences are neighbors in sequence space), and the mapping from sequence space to a phenotype. In order to determine the fitness landscape for a particular lattice and chain length, all sequences that fulfill the criteria for a stable, functional sequence must be found. The concept was introduced by Wright[10] in 1930 and remains a topic of current interest.[11] It is a central idea about how molecular evolution occurs.

Other models of fitness landscapes include the *NK* model[12] and the "block *NK*" model.[13,14] 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 form of the model. Each element is randomly assigned a 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, until $K=N-1$, where the landscape is uncorrelated.

The effects of the requirements of protein folding and function can be addressed by modeling protein structure explicitly with molecular-dynamics simulations. Using such a model, Yomo *et al.*[15] 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. Lattice models are

[a)]Electronic mail: jonathan.hirst@nottingham.ac.uk an attractive alternative for generating a genotype-phenotype mapping, as they are both physically relevant and computationally tractable.[4] A population dynamics simulation of lattice model proteins gave results consistent with the molecular-dynamics simulation of Yomo *et al.*[15] in that evolution of a lattice model protein for function resulted in a population of stable protein sequences.[8]

In this study, we characterize the fitness landscape described in a previous work[6] on HP chains[16] on a diamond lattice through deterministic computer simulations of the population. The fitness landscape is limited to those sequences that are minimally fit (i.e., fold to a stable structure and possess a binding pocket). We use the shifted-HP model[17] which includes a repulsive energy between polar residues. Residues can be either hydrophobic (H) or polar (P). The interaction energy between two H residues is $E_{HH}=-2$. The energy for all other interactions is given by $E_{HP}=E_{PH}=E_{PP}=+1$. In order to be deemed a native structure for a given sequence, the energy of that conformation must be lower by two units than any other possible structure. This minimizes the possibility of classing sequences with uncooperative folding as stable, without resorting to intensive simulations of the folding kinetics.[6] The inclusion of repulsive interactions results in native structures that are not necessarily maximally compact and so include one or more unoccupied lattice sites that can be characterized as binding pockets. Previous studies have examined these sequences on the square[18,19] and diamond[6] lattices. The definition of fitness means that the whole of sequence space for length $n=25$, for example, will consist of a set of disconnected networks with each isolated network comprising fit sequences connected by single point mutations. A lethal mutation is equivalent to stepping off the fitness landscape, and nonfit sequences are omitted in the figures.

In this paper we investigate generic features of protein evolution by population dynamics simulations on a fitness landscape arising from a simple, but explicit, physical model of proteins. We study whether it is possible to rationalize the observed evolutionary behavior from analysis of the nature of the fitness landscape, without resorting to population dynamics. We develop a model of population dynamics on a fitness landscape based on an analogy with a thermodynamic system. A fitness landscape can be considered to have a set of states, each with a given fitness (i.e., energy). We address several questions that arise from use of the model. Does the nature of fitness landscapes allow population dynamics to proceed in a broadly thermodynamiclike manner? How does the nature of fitness landscapes affect the final populations? Do some fitness landscapes exhibit features that drive the final population away from what would be expected from the thermodynamic analogy of the system?

II. METHODS

A. Modeling evolution

We adapt a model previously used to examine evolution with recombination.[7] The progress of evolution is governed by Eq. (1), which relates the population $p_i(q+1)$ of a viable sequence i in generation $q+1$ to the populations of generation q,

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

Here, μ is the point mutation rate, 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 increases the population of sequence i by $(\mu/n)p_j(q)$ through mutation. The factor $\mathcal{N}(q)$ normalizes the populations so that

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

In each generation, the effect of fitness is accounted for by multiplying the population by a factor f_i the fitness of sequence i. Previously[7] it has been taken to be a function of stability. Here, we define the fitness based on the number of hydrophobes h_i in the nearest and next-nearest neighbors of the binding pocket, which was introduced as a measure of fitness related to the nonspecific binding of a hydrophobic substrate in previous work,[6]

$$f_i=e^{\alpha h_i}. \tag{2}$$

There are only four nearest neighbors on the diamond lattice, which does not allow much functional diversity (but is obviously advantageous for enumerating structures). So the eight next-nearest neighbors are also considered, i.e., potentially 12 residues of the 25-mer determine its fitness. We incorporate into the definition of fitness a selection coefficient (α), which is analogous to an inverse temperature. At low α, evolution becomes more neutral; conversely at high α, selection will dominate when possible.

The simulations can be treated as a set of matrix operations on a population vector. We begin our population dynamics experiments by "seeding" each simulation with a population of unity for 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. Many families of related sequences were found from our study on the diamond lattice.[6] Of these families, we examine the largest 25-mer family of 151 sequences (Fig. 1), the second largest family of 120 sequences (Fig. 2), and the 86 sequence family (Fig. 3), which is notable for the presence of a bottleneck and kinetic trap. Typical structures of proteins in the three families are shown in Fig. 4. These proteins are stable because the criteria for a viable functional model protein require an energy gap between the unique native conformation and the first excited-state conformation.

Each family is made up of several neutral nets, sets of interconnected sequences that share the same fitness. 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

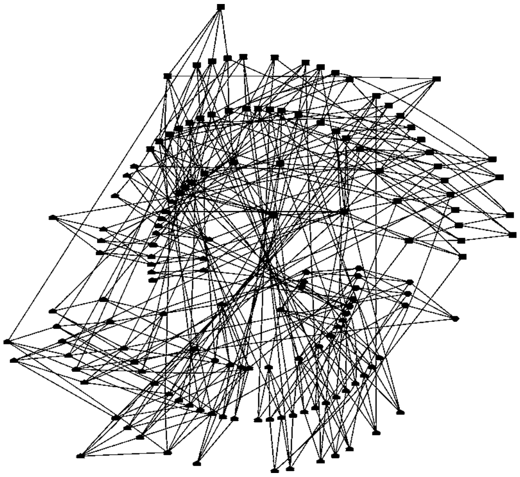


FIG. 1. 151-member fitness landscape of three-dimensional 25-mer functional model proteins. The nodes correspond to functional model proteins; the edges correspond to single point mutations connecting two viable sequences. The nodes closest to the center are "hub nodes." The node shape indicates the fitness (the number of H residues in the binding pocket and the layer surrounding the binding pocket). Circular nodes are of fitness 3, triangular=4, rectangular=5, pentagonal=6, and hexagonal=7. While this figure only contains nodes with fitness 4, 5, and 6, other figures contain the full range.

able to make all these connections individually. Furthermore, these nodes often connect to each other. We refer to these nodes as "hubs," and they are plotted at the center of Figs. 1–3. Our families conform to a "superfunnel" structure,[20,21] with the hub nodes the most stable sequences and the stability decreasing towards the rim of the landscape.

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,

$$\delta_p(q) = \sum_{i=1}^{C} (p_i(q) - p_i(q-1))^2. \tag{3}$$

A population dynamics simulation terminates when $\delta_p < 10^{-20}$. Fluctuations below this level are insignificant, unless we are concerned with very small changes in large ($>10^{10}$) populations.

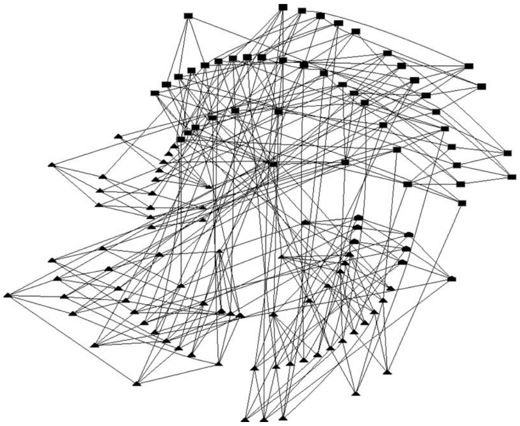


FIG. 2. 120-member fitness landscape of three-dimensional 25-mer functional model proteins, drawn with the same convention as Fig. 1.

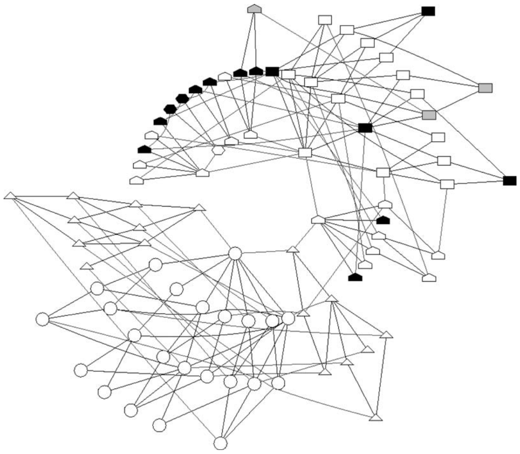


FIG. 3. 86-member fitness landscape of 25-mers, drawn with the same convention as Fig. 1. The node color indicates the time taken to reach the steady state at α=0.3 and μ=0.01. White indicates ≈5500 generations, grey ≈6700 generations, and black ≈12500 generations.

To draw an analogy with thermodynamics, a low selective pressure will result in a population governed by entropy rather than energy. However, entropy (and enthalpy) on fitness landscapes cannot be directly connected to free energy, as the process of population dynamics is not identical to that of chemical kinetics. For this reason, we use landscape entropy and landscape enthalpy when referring to processes on the fitness landscape. Landscape entropy is different to a sum of the number of nodes; it also involves the connectivity of the graph.

III. THERMODYNAMIC ANALOGY

If we consider fitness as a landscape enthalpy, then we may consider α, μ, and n to be components of a "temperature." For simplified fitness landscapes we can derive a function to compute the steady-state populations from the values of α, μ, and n.[21] The complexity of this solution for even a simplified fitness landscape suggests that any thermodynamic analogy will be an approximation.[21] If, in practice, populations on the fitness landscapes of functional model proteins do behave in a manner broadly akin to chemical thermodynamics, this tells us about the nature of the fitness landscapes.

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

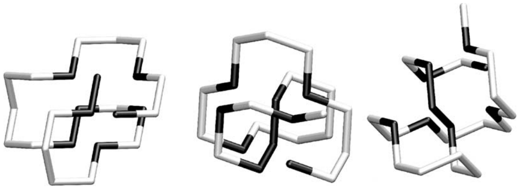


FIG. 4. Representative examples of functional model proteins on the diamond lattice. White residues are polar, black are hydrophobic. Examples are drawn from the 86-member (left), 120-member (middle), and 151-member (right) families examined in this study.

$$p_i = \frac{e^{-\beta \varepsilon_i}}{Q}, \tag{4}$$

where

$$Q = \sum_j e^{-\beta \varepsilon_j}. \tag{5}$$

If $\beta=0$, all states would be equally populated. This is clearly not the case with our population dynamics simulations where, due to the connectivity of the fitness landscape, there will be different occupancies even when the selection coefficient $\alpha=0$ (i.e., when $\beta=0$). This is analogous to thermodynamics, where entropy alters 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

$$p_i = w_i \frac{e^{-\beta \varepsilon_i}}{Q}. \tag{6}$$

We accommodate the effect of connectivity on the fitness 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, λ_i, of next-nearest neighbors for sequence i:

$$p_i = \sigma(\lambda) \frac{e^{-\beta \varepsilon_i}}{Q} = \frac{e^{-\beta \varepsilon_i + \ln \sigma(\lambda_i)}}{Q}. \tag{7}$$

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

A. Landscape enthalpy and landscape entropy

We equate the landscape enthalpy of a sequence (or set of sequences) with its fitness,

$$\varepsilon_i = -h_i. \tag{8}$$

There is no direct relationship between the number of states of a molecule and the connectivity of the fitness landscape around a sequence, which would lead to a strict definition of landscape entropy. A relationship between physical systems and fitness landscapes can be drawn, however, from consideration of the population at high temperature, where enthalpy becomes a negligible consideration and the landscape entropy dominates. In the same way, we can consider the entropy of a sequence to be related to its steady-state population when $\alpha=0$, i.e., under threshold selection.

We define landscape entropy from the local connectivity of the fitness landscape, based on the observation that there is a high correlation between the number of next-nearest neighbors and the population of sequences at steady state under threshold selection.[21] 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 by mutations. For threshold selection the only way to gain population is by renormalization (which cannot increase the population of one sequence over another) or through receiving population through mutations from neighboring sequences. Thus, there is some correlation between the number of nearest neighbors and population under threshold selection. However, the amount of population donated by the neighbors depends on their own populations, which in turn depends on their own neighbors. A good approximation to this effect is a count of next-nearest neighbors. On larger fitness landscapes, it may be advantageous to extend this to further reaches of the landscape.

The correlation between next-nearest neighbors and landscape entropy depends on the population of each node not deviating too sharply from the mean; i.e., all nodes are considered equal. The landscape entropy is calculated from all next-nearest neighbors, 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 landscape entropy as the total number of next-nearest neighbors for all sequences of a given fitness.

The relationship between the population of sequence i and the number of its next-nearest neighbors λ_i was found[21] to be linear when $\alpha=0$, i.e., of the form:

$$p_i = \mathcal{K} \lambda_i + \mathcal{C}. \tag{9}$$

Here, $\mathcal{K}$ and $\mathcal{C}$ are constants. Since $\mathcal{C}$ is small, and for the majority of sequences λ_i is large, we neglect $\mathcal{C}$. As already outlined, we consider whole populations of sequences with identical fitnesses:

$$p_s = \mathcal{K} \sum_{i=0}^{c_s} \lambda_i. \tag{10}$$

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

$$p_s = \frac{\sum_{i=0}^{c_s} \lambda_i}{\sum_{j=0}^{c} \lambda_j}. \tag{11}$$

From the partition function [Eq. (7)]:

$$\frac{\sum_{i=0}^{c_s} \lambda_i}{\sum_{j=0}^{c} \lambda_j} = \frac{\sigma(\lambda_s) e^{-\beta \varepsilon_s}}{Q_{\alpha=0}}. \tag{12}$$

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

$$\frac{\sum_{i=0}^{c_i} \lambda_i}{\sum_{j=0}^{c} \lambda_j} = \frac{\sigma(\lambda_s)}{Q_{\alpha=0}}. \tag{13}$$

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

$$\sigma'(\lambda_i) = \frac{\sum_{s=0}^{c_s} \lambda_i}{\sum_{j=0}^{c} \lambda_j}. \tag{14}$$

Hence,

$$\ln \sigma(\lambda_s) = \ln \sigma'(\lambda_s) + \ln Q_{\alpha=0}. \tag{15}$$

The inverse temperature β is a function of α, μ, and n, and is determined empirically on the 151-member landscape. Details are given in the Appendix. In this section we have defined a landscape enthalpy and landscape entropy. In Sec. IV we assess the utility of these quantities for estimating populations from a landscape.

IV. RESULTS

This paper attempts to probe why evolution on some fitness landscapes is well predicted by the partition function, rather than to derive a general function that is applicable in all cases. Therefore, for clarity, only variation of the selection coefficient α is considered further; the point mutation rate μ is kept fixed. We use β_α (see the Appendix) to make predictions of the populations for several different fitness landscapes under various conditions. Figure 5 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, the method captures the trends quantitatively for the other two fitnesses which were not explicitly included in the development of the β function.

Similar results were obtained with the fitness landscape of 120 sequences, shown in Fig. 2. A fitness landscape with a much different structure is the 86-member landscape (Fig. 3). This is structured as two sublandscapes, as discussed previously,[6] which makes populations on this fitness landscape difficult to estimate with the thermodynamic analogy. The estimated population curves are qualitatively, although not quantitatively, correct (Fig. 6).

The difference between estimated and real populations on the 86-member fitness landscape arises from the breakdown of the assumption described earlier that the number of next-nearest neighbors is a good indication of population at the steady state under threshold selection. This is not the case for this fitness landscape, which leads to the discrepancies seen in Fig. 6 at low values of the selection coefficient. These arise from the nature of the fitness landscape; there are effectively two landscapes in competition with each other. Alteration of the structure of one sublandscape will appear to have minimal direct effects on the entropic factors in the other sublandscape and will not be captured by the landscape-entropy function. However, because populations on those sublandscapes are still in direct competition with each other, a strong effect is observed.

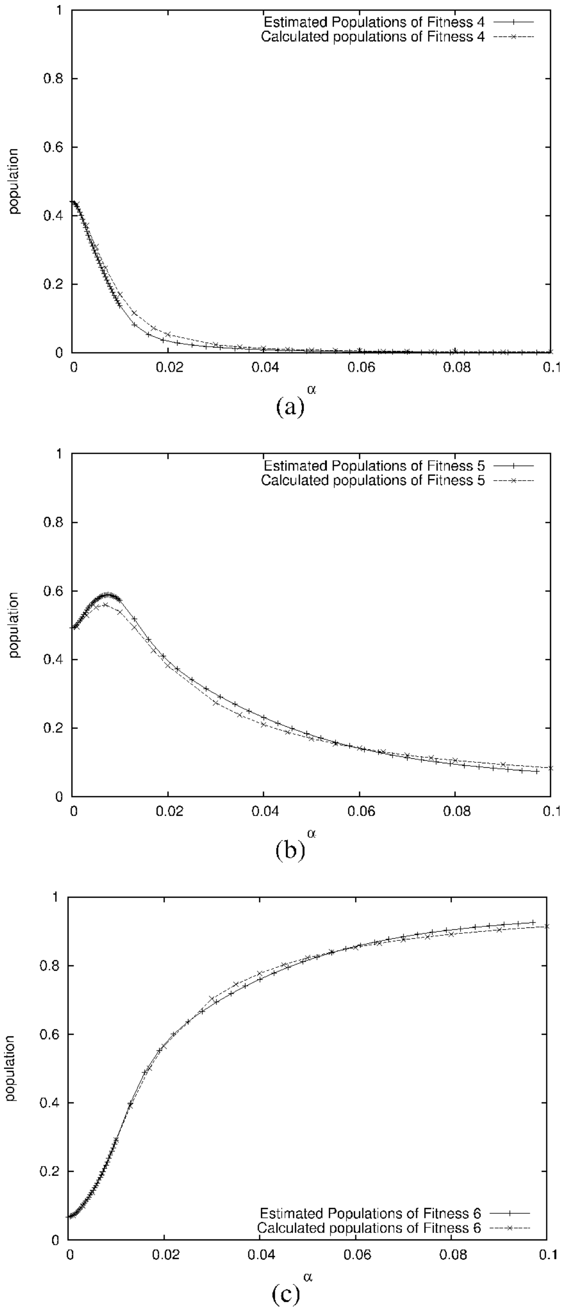
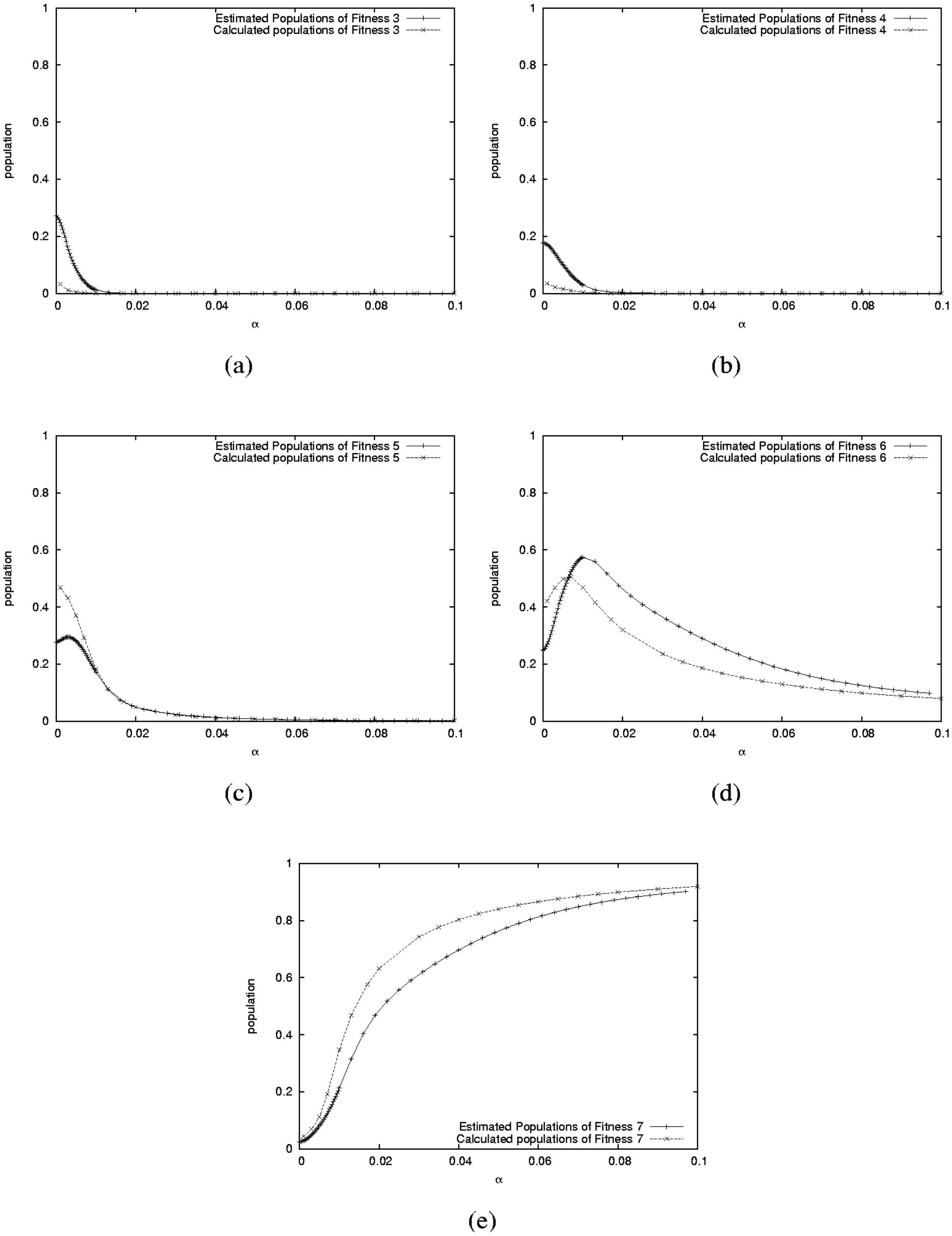


FIG. 5. Populations of (a) fitness 4, (b) fitness 5, and (c) fitness 6 for the 151-member fitness landscape (shown in Fig. 1). Populations are calculated through the population dynamics method and estimated using the analogy to thermodynamics described in the text.

A. Rate of convergence under neutral evolution

We observe that, under the population dynamics scheme given earlier, different arbitrary initial populations converge 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 steady state is found by running a simulation until the sum of the squares of the differences in population from one generation to another [Eq. (3)] is less than 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 fittest area of the fitness 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 1 and the time taken to converge to the steady state is recorded. The results for the 86-sequence family (Fig. 3) show a partitioning of sequences into those which take ≈5500, ≈6700, and



(a)

(b)

(c)

(d)

(e)

FIG. 6. Steady-state populations of (a) fitness 3, (b) fitness 4, (c) fitness 5, (d) fitness 6, and (e) fitness 7 for the 86-member fitness landscape (shown in Fig. 3). Populations are calculated through the population dynamics method and estimated using the analogy to thermodynamics described in the text.

≈12 500 generations to reach the steady state. As soon as a sequence of fitness 7 is populated it easily outcompetes the other sequences and its population rapidly increases. There is then neutral evolution among the three sequences (of fitness 7) 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 Fig. 7 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 (of fitness 7) 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. In the slowest case, 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 neighboring sequences contribute to both the nonhub peak sequences equally.

For the 151-member fitness landscape the situation is more complex, although the sequences can again be partitioned into different sets based upon their rates of convergence. The largest group of 75 sequences take ≈7400 generations to converge. The remaining sequences can be grouped into five sets of 5–15 sequences and 25 small groups of 1–3 sequences. This fitness landscape is more variable than the 86-member landscape, but still around half (75) of the 151 sequences undergo neutral evolution to converge at the same rate.

Neutral evolution is an important component of molecu-

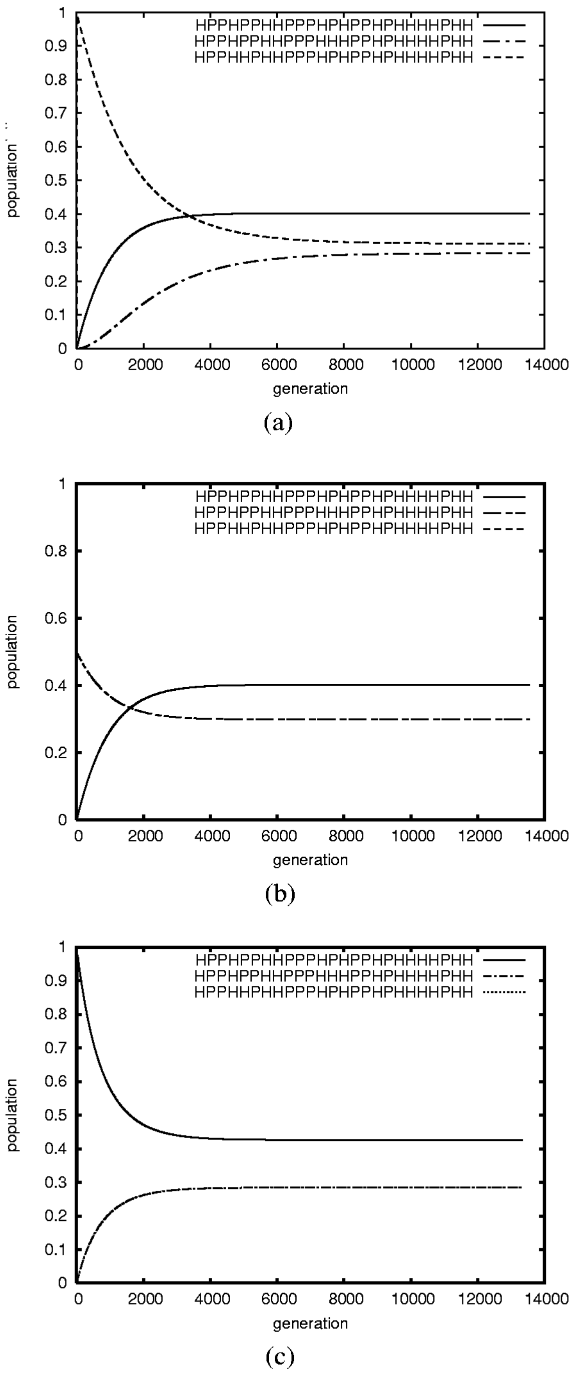


FIG. 7. Population of each sequence of fitness 7 over the course of a population dynamics simulation, for starting sequences leading to (a) slow, (b) medium, and (c) fast convergence. In (b) and (c) the populations of the two sequences represented by the dashed lines are identical.

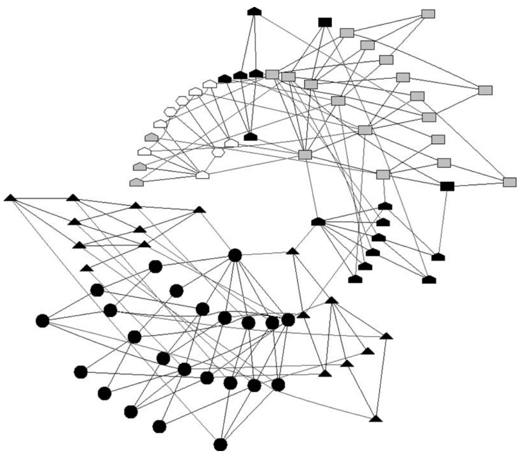


FIG. 8. 86-member fitness landscape of 25-mers. The color and shape indicate fitness as in Fig. 1. The node color indicates the number of generations before 99% of the population occupies the peak (hexagonal nodes). White indicates 0–39 generations, grey 40–79 generations, and black 80–124 generations.

lar evolution.[22] Neutral drift may be important as a necessary step before new function can be acquired.[23] Here, 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 of maximum fitness. Under some fitness landscapes this may result in a shorter or longer period of neutral drift being necessary before a 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 effect of neutral drift.

B. Rate of convergence under adaptive evolution

The populations rapidly reach the fittest sequences. In the previous set of population dynamics simulations, most of the time 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 of maximum fitness, 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 8 shows this for the 86-member fitness landscape, with $\alpha=0.3$ and $\mu=0.01$. The numbers of generations shown in Fig. 8 are about two orders of magnitude less than those in Fig. 3. On the fitness landscapes, the bulk of adaptive evolution takes place rapidly, followed by a long period of largely neutral evolution, before reaching a steady state.

We observe that there are three "kinetic traps" that can slow evolution in the 86-member fitness landscape. These are neutral nets with fitness 6 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 3 and 4 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 3 and 4 is similar to that of fitness 6. The region of fitness 6 acts as a bottleneck, and so the time taken to evolve

TABLE I. Effect of movement of the seeding sequence towards the hub nodes on the rate of adaptive evolution on the landscape shown in Fig. 1.

Effect of mutation on		
Rate of adaptive evolution	Distance to hub nodes	Number of mutations
Increase	Closer	189
Decrease	Closer	24
Any	None	107
None	Any	106
None	None	9

a significant population of fitness 6 is negligible in comparison to the time taken to move across to the global maximum of fitness 7.

The fitness landscapes display a superfunnel[20] topology. The nodes that are more thermodynamically stable (with respect to unfolding) are closer to the center 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. Table I shows that most mutations that increase the rate of adaptive evolution involve a move in the direction of the hub nodes (which are more thermodynamically stable with respect to folding). Thus, increasing the thermodynamic stability of a sequence will likely result in a more rapid adaptive evolution within a superfunnel.

V. CONCLUSIONS

We have considered how the nature of some fitness landscapes allows population dynamics to proceed in a manner akin to chemical thermodynamics, allowing estimation of steady-state populations from a partition function. The landscape entropy in this process is not the number of sequences of a particular fitness, but rather the connectivity of the fitness landscape. We observe fitness landscapes which are modeled quite accurately by the partition function. However, in some fitness landscapes the connectivity lowers the quantitative accuracy of predictions from this scheme, although the trends of the population changes under changing selection pressures are still qualitatively accurate. Why might predictions be qualitatively good but quantitatively poor? For the instances given here, it seems that the division of a fitness landscape into two weakly linked sublandscapes is responsible. This increases the relative population of one part of the fitness landscape beyond what the partition function is able to predict. This may be a consequence of the formulation of the population dynamics process presented here, in which a small increase in the number of viable sequences or the mean number of neighbors of a sequence results in a large increase in mutational robustness, with a corresponding larger relative population for that area of the fitness landscape.

An assumption of the model is that the population is infinite and so, given that the fitness landscape does not contain any disconnected regions, simulations will always converge to the same point. This allows us to measure the progress of the simulation in terms of a known end point, but certain aspects of the fitness landscape, such as possible "kinetic traps," cannot be fully analyzed.

We have focused on point mutations. A few other studies have considered recombination mutations. The effect of the allowed transitions in sequence space has been tested with a lattice model for genotype-phenotype mapping.[7] 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[24] and found that a restricted ratio of recombination to point mutations is necessary to find an optimal sequence on a flat fitness landscape.

Although we do not use the thermodynamic analogy to predict populations of individual sequences, it is straightforward 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 neighbors are counted. These sequences can be neglected because they account for only a tiny proportion of the donated population onto the peak in the steady state. For populations under weak selection, the thermodynamic analogy predicts individual sequences well, as long as the fitness landscape is not splitted into weakly connected sublandscapes. For sequences under intermediate selection, the process works well for averaged populations, but less so for individual sequences.

The superfunnel structure of the fitness landscapes leads to the hub nodes being favored by evolution and thereby the over-representation of these sequences during the course of evolution. However, the sheer number of nodes near the rim of the fitness landscape causes the hub node populations to be low. On larger, more realistic fitness 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,[25–27] could benefit from further understanding of the nature of fitness 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 center of the funnel and so a greater number of acceptable mutations become available to them. One argument (as yet unproven) against such a strategy might be that the marginal stability of proteins is a required part of their function.[28]

Does natural evolution utilize superfunnels to improve 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 fitness landscapes for longer sequences are likely to be more extreme than the landscapes seen here. However, these studies do not address changes in mutation rate, for example, by bacteria under the "SOS response."[29] It may be that a whole series of neutral mutations is necessary to move the population towards the hub before mutation to a 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. 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[30] on our fitness landscapes could lead to some insights into this possible process.

ACKNOWLEDGMENTS

We thank Andrew Gilbert for useful discussions. We thank the BBSRC Bioinformatics Committee for a Ph.D. studentship and the EPSRC for the computing equipment (GR/62052).

APPENDIX

Below we determine β, the inverse temperature. We can rewrite the partition function in Eq. (7) in terms of all sequences of a certain fitness, rather than individual sequences:

$$p_s = \frac{e^{-\beta\varepsilon_s + \ln\sigma(\lambda_s)}}{Q}, \tag{A1}$$

where p_s is the predicted total population of all sequences of set s and ε_s is the landscape enthalpy of sequences in set s. We combine Eqs. (8), (15), and (A1) below:

$$p_s = \frac{e^{\beta h_s + \ln\sigma'(\lambda_s)}}{\sum_z e^{\beta h_z + \ln\sigma'(\lambda_z)}}. \tag{A2}$$

The inverse temperature 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 be populated by mutations from other sequences. Each generation, μ/n of the population is donated to each neighboring 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 paper we exclusively consider fitness landscapes where $n=25$, we neglect it as a factor.

We characterize the relationship between μ, α, and β empirically. We vary α from 0.0005 to 1, keeping μ constant at 0.1, and run a population dynamics simulation on the largest, 151-member fitness landscape (Fig. 1) until the steady state is reached. We find a value of β by solving Eq. (A2) with p_s taken as the population of the set of sequences with fitness 6. We repeat this process for values of μ from 0.001 to 0.9, keeping α constant at 0.1.

The empirically derived functions for μ and α are given below, where β_α is the fit to α and β_μ the fit to μ. These functions are plotted in Fig. 9.

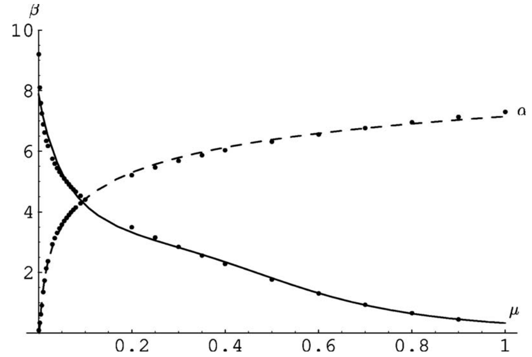


FIG. 9. The relationship between β (i.e., an "inverse temperature" function) and μ (solid line) and between β and α (dashed line). The functions are fitted with the equations in the Appendix.

$$\beta_\alpha = 17.3535 + \frac{0.041\,020\,4}{\alpha^{0.670\,229}} - \frac{10.1479}{\alpha^{0.107\,302}}, \tag{A3}$$

$$\beta_\mu = \frac{1}{0.126\,652 + 1.4264\mu - 3.841\,719\mu^2 + 5.372\,44\mu^3}. \tag{A4}$$

[1] Y. Xia and M. Levitt, Curr. Opin. Struct. Biol. **14**, 202 (2004).
[2] H. S. Chan and E. Bornberg-Bauer, Appl. Bioinf. **1**, 121 (2002).
[3] N. Go, Int. J. Pept. Protein Res. **7**, 313 (1975).
[4] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, Protein Sci. **4**, 561 (1995).
[5] E. Bornberg-Bauer, Z. Phys. Chem. **216**, 139 (2002).
[6] B. P. Blackburne and J. D. Hirst, J. Chem. Phys. **119**, 3453 (2003).
[7] Y. Cui, W. H. Wong, E. Bornberg-Bauer, and H. S. Chan, Proc. Natl. Acad. Sci. U.S.A. **99**, 809 (2002).
[8] P. D. Williams, D. D. Pollock, and R. A. Goldstein, J. Mol. Graphics Modell. **19**, 150 (2001).
[9] G. Tiana, N. V. Dokholyan, R. A. Broglia, and E. I. Shakhnovich, J. Chem. Phys. **121**, 2381 (2004).
[10] S. Wright, Proceedings of the Sixth International Congress on Genetics, Brooklyn Botanic Garden, Menasha, WI, 1932 (unpublished), p. 354.
[11] E. Lieberman, C. Hauert, and M. A. Nowak, Nature (London) **433**, 312 (2005).
[12] S. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford University Press, Oxford, 1993).
[13] A. S. Perelson and C. A. Macken, Proc. Natl. Acad. Sci. U.S.A. **92**, 9657 (1995).
[14] L. D. Bogarad and M. W. Deem, Proc. Natl. Acad. Sci. U.S.A. **96**, 2591 (1999).
[15] T. Yomo, S. Saito, and M. Sasai, Nat. Struct. Biol. **6**, 743 (1999).
[16] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[17] H. S. Chan and K. A. Dill, Proteins **24**, 335 (1996).
[18] J. D. Hirst, Protein Eng. **12**, 721 (1999).
[19] B. P. Blackburne and J. D. Hirst, J. Chem. Phys. **115**, 1935 (2001).
[20] E. Bornberg-Bauer and H. S. Chan, Proc. Natl. Acad. Sci. U.S.A. **96**, 10689 (1999).
[21] B. P. Blackburne, Ph.D. thesis, The University of Nottingham, 2004.
[22] M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1983).
[23] D. J. Lipman and W. J. Wilbur, Proc. R. Soc. London, Ser. B **245**, 7 (1991).
[24] Y. Xia and M. Levitt, Proc. Natl. Acad. Sci. U.S.A. **99**, 10382 (2002).
[25] P. Kast and D. Hilvert, Curr. Opin. Struct. Biol. **7**, 470 (1997).
[26] Z. X. Shao and F. H. Arnold, Curr. Opin. Struct. Biol. **6**, 513 (1996).
[27] J. Minshull and W. P. C. Stemmer, Curr. Opin. Chem. Biol. **3**, 284 (1999).
[28] D. M. Taverna and R. A. Goldstein, Proteins **46**, 105 (2002).
[29] H. Echols and M. F. Goodman, Mutat Res. **236**, 301 (1990).
[30] T. B. Kepler and A. S. Perelson, Proc. Natl. Acad. Sci. U.S.A. **92**, 8219 (1995).