Neutral evolution of proteins: The superfunnel in sequence space and its relation to mutational robustness.

Josselin Noirel, Thomas Simonson

Neutral evolution of proteins: The superfunnel in sequence space and its relation to mutational robustness

Josselin Noirel[a] and Thomas Simonson[b]
Laboratoire de Biochimie, École Polytechnique, Route de Saclay, Palaiseau 91128 Cedex, France

(Received 31 July 2008; accepted 11 September 2008; published online 11 November 2008)

Following Kimura's neutral theory of molecular evolution [M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1983) (reprinted in 1986)], it has become common to assume that the vast majority of viable mutations of a gene confer little or no functional advantage. Yet, *in silico* models of protein evolution have shown that mutational robustness of sequences could be selected for, even in the context of neutral evolution. The evolution of a biological population can be seen as a diffusion on the network of viable sequences. This network is called a "neutral network." Depending on the mutation rate μ and the population size N, the biological population can evolve purely randomly ($\mu N \ll 1$) or it can evolve in such a way as to select for sequences of higher mutational robustness ($\mu N \gg 1$). The stringency of the selection depends not only on the product μN but also on the exact topology of the neutral network, the special arrangement of which was named "superfunnel." Even though the relation between mutation rate, population size, and selection was thoroughly investigated, a study of the salient topological features of the superfunnel that could affect the strength of the selection was wanting. This question is addressed in this study. We use two different models of proteins: on lattice and off lattice. We compare neutral networks computed using these models to random networks. From this, we identify two important factors of the topology that determine the stringency of the selection for mutationally robust sequences. First, the presence of highly connected nodes ("hubs") in the network increases the selection for mutationally robust sequences. Second, the stringency of the selection increases when the correlation between a sequence's mutational robustness and its neighbors' increases. The latter finding relates a global characteristic of the neutral network to a local one, which is attainable through experiments or molecular modeling. © *2008 American Institute of Physics*. [DOI: 10.1063/1.2992853]

I. INTRODUCTION

In silico simulations have provided in-depth insights into the underlying mechanisms of molecular evolution.[1,2] These models rely on the physical properties of simplified protein or RNA molecules, making use of the thermodynamic stability in most cases and allow the researcher to examine a population's genetic history in the greatest detail. Many aspects of evolution can be simulated: from the scale of the gene or protein[3–5] to the scale of the genome,[6] and from the rate of substitution[7–9] to the link between functionality and stability.[5,10–12]

Maynard Smith introduced the concept of "protein space" in an attempt to understand how evolution proceeded.[13] Evolution occurs through mutation, and this prompts us to investigate which sequences may perform a given function (for instance, catalyzing a given enzymatic reaction) and which of them are reachable via mutations. Lipman and Wilbur addressed this question using a lattice model of a protein where each residue corresponds to one bead on the lattice and may be either hydrophilic or polar (HP model).[14–16] Though simple, the HP model led to the first observation that different isoalleles corresponding to a same phenotype were clustered in sequence space and connected via mutations. These clusters were named "evolutionary networks" and will be hereafter referred to as "neutral networks." "Neutral" refers to the absence of intrinsic selective advantage of any sequence as first formulated by Kimura:[17] all viable sequences perform equally well. The nature and structure of the neutral networks were further investigated and it was discovered that, for a given phenotype, viable sequences were organized into a "superfunnel."[1,5,18–22] This was achieved using *in silico* models and thereafter confirmed experimentally.[10,23] Briefly, the superfunnel topology consists of a core of alleles, "the prototype sequences," these are sequences that encode thermodynamically stable proteins and that are highly robust to mutations. By "sequence robust to mutations," we mean that mutations are unlikely to disrupt the structure and function of the protein encoded by the sequence. In other words, the mutation of a robust sequence is more likely to produce a sequence which still belongs to the neutral network. The core is surrounded by a large halo of alleles which are less robust to mutations and encode less stable proteins.

How the sequences are organized is not only interesting by itself but it also determines the usage of the sequences within a population of individuals. In other words, the network topology affects the long-term probability for each se-

[a]Electronic mail: j.noirel@sheffield.ac.uk.
[b]Electronic mail: thomas.simonson@polytechnique.edu.

quence to appear within the population.[19] Two limiting regimes exist depending on the total mutation rate. When the total mutation rate is small, evolution essentially resembles a random walk on the network and all the viable sequences are equiprobable. However, when the total mutation rate is large, evolution favors the alleles that are more robust to mutations; under such circumstances, the stability is indirectly selected for by reason of the superfunnel topology. More rigorously, the transition occurs when $\mu N \approx 1$, where μ is the mutation rate (per gene and per generation) and N is the population size.[9,24–26]

It was rapidly realized that, alongside the parameters μ and N, the superfunnel topology determined the statistical properties of evolution. Nonetheless, little work was devoted to investigating the extent to which the topology could skew the probability of a particular sequence.[18] This is the question that we address in this study through the use of two models: a two-dimensional (2D) on-lattice and a three-dimensional (3D) off-lattice model. We first compare the results from the two models and demonstrate that both models give rise to the same qualitative degree distribution for a neutral network. This suggests that the superfunnel topology does not depend strongly on the details of the structural model. We then compare typical neutral networks to null models of networks, namely, Erdős–Rényi random networks and random scale-free networks, to help pinpoint the topological features that are important in increasing the selection for mutationally robust sequences. Finally, by generating random networks with the same degree distribution as a neutral network, we show that the extent of selection for robust sequences is highly correlated with the smoothness of the mutational robustness across the network. This observation allows us to relate a global property of the network, the selection for mutational robustness, to a local one, and the smoothness of the mutational robustness across the network.

II. MATERIALS AND METHODS

A. Structural models

A neutral network is a network where the nodes are the sequences able to perform a vital function. The edges of the network connect pairs of sequences which differ from each other at only one position. We shall assume that, for a protein sequence, to perform the vital function is equivalent to folding stably into a given conformation. The rationale behind the choice of foldability is the observation that preserving the three-dimensional (3D) structure is a necessary condition for an enzyme to be functional. To evaluate the foldability and the thermodynamic stability, it requires a structural model of the protein. Sections II A 1 and II A 2 introduce two such models.

1. Two-dimensional on-lattice model

The first model and its variations were employed by many researchers in the past.[4,5,25,27] In this model, the proteins are made up of 25 amino acids, which can be either hydrophobic (H) or polar (P). The protein chain is considered to have a direction; for instance, the sequences HPP and PPH are different. The amino acids are treated as beads and the conformations are restricted to a 5×5 lattice. Thus, conformations that are not maximally compact are discarded. There are 1081 such conformations that are unrelated by symmetry. The energy of a sequence s in a particular conformation c is given by

$$E(s,c) = \sum_{i<j} e_{ij}\Delta_{ij}, \tag{1}$$

where Δ_{ij} is equal to 1 if the ith and jth amino acids are in contact, that is, if they are neighbors in the grid but not within the sequence and to zero otherwise; furthermore, $e_{ij}=e_{\mathrm{HH}}$ if both amino acids i and j are hydrophobic, $e_{ij}=e_{\mathrm{PP}}$ if both amino acids are polar, and $e_{ij}=e_{\mathrm{HP}}$ otherwise. Following Li *et al.*,[27] the following values were used:

$$e_{\mathrm{HH}} = -2.3, \quad e_{\mathrm{HP}} = -1.0, \quad e_{\mathrm{PP}} = 0.0. \tag{2}$$

These values ensure that the proteins display a hydrophobic core and a preferentially hydrophilic surface. The native conformation of a sequence s is the conformation c which minimizes the energy $E(s,c)$ and such a conformation must be unique with respect to s. In other words, for a sequence to fold, its ground state has to be nondegenerate.

If one focuses one's attention onto a particular conformation c which is supposed to carry out a vital function, a neutral network can be constructed by connecting any pair of sequences that fold into c and which differ from each other at only one position in their amino acid sequence. There might exist several connected network components. However, a giant component generally gathers the vast majority of the viable sequences. Furthermore, the smallest components are of little interest, since the sequences cannot "jump" from one component onto another. We therefore concentrate, without loss of generality, on the giant component of each neutral network. The mutational robustness of a sequence is equal to the number n of neighbors this sequence possesses within the network. Additionally, the thermodynamic stability of a sequence is evaluated by computing the folding temperature T_f, which is the temperature at which the Boltzmann probability of the native conformation is equal to 0.5.

2. Three-dimensional off-lattice model

The second structural model is a three-dimensional (3D) off-lattice protein description.[5,28,29] Three proteins are considered: the 69-residue SH3 domain of Vav, the 57-residue SH3 domain of Grb2 (PDB accession number: 1gcq), and the 20-residue TRP-cage (PDB accession number: 1l2y). The TRP-cage peptide, thanks to its very small size, allows more thorough computations. For each of these three structures, a set of decoy structures is prepared: 1135 for Vav; 1285 for Grb2; 1791 for TRP-cage. These decoy structures are prepared by threading the amino-acid sequences through unrelated protein structures selected from the Protein Data Bank.[30] Furthermore, 100 additional decoy structures are prepared for Vav and Grb2 by means of molecular dynamics *in vacuo* at 310 K using CHARMM.[31] The latter structures possess more nativelike contacts.

The energy of a sequence s in a conformation c is given by the same equation (1) as above. However, there are important differences with respect to the on-lattice model. First,

the elementary terms e_{ij} may be taken from a more complex energy matrix. Depending on the degree of accuracy desired, the amino acids may be classified into 2, 3, 4, 6, or 20 categories, at which point the maximum degree of resolution with this model is reached. These various classifications are called "folding alphabets."[32] Given an alphabet, the value e_{ij} only depends on the classes of the amino acids i and j. The values $\mathcal{E}=(e_{ij})$ form the "energy matrix." The alphabets, along with the associated matrices, were optimized earlier.[28] Second, a contact between the ith and jth amino acids occurs $(\Delta_{ij}=1)$ if at least one of i's heavy atoms is within 4.5 Å from one of j's. The positions of the atoms are computed by grafting the most frequent rotamer onto the backbone.[5,28,33] We discard any contact occurring between two amino acids which are separated by fewer than two residues along the protein sequence. When a two-class alphabet is used the optimization converges toward the following set of parameters:[28]

$$e_{\mathrm{HH}}=-8.5, \quad e_{\mathrm{HP}}=+9.0, \quad e_{\mathrm{PP}}=-3.5 \tag{3}$$

corresponding to the following classification: LVIMCASTPGFWY are considered hydrophobic and EDNQKRH are considered polar.

For a particular protein, the sequence s is said to fold into the native conformation c if the following conditions are met:

(1) the energy of the sequence in the native conformation is lower than in any decoy conformation,
(2) the energy gap ΔE between the native conformation and any decoy conformation must be greater than a threshold ΔE_0, and
(3) the Z-score of the native conformation must be lower than a threshold $Z_0<0$.

The values ΔE_0 and Z_0 are evaluated by computing the energy gap and the Z-score for a sequence derived from the native sequence by a short minimization (in sequence space). This minimized sequence will be referred to as the "reference sequence." Minimization is achieved by mutating randomly and iteratively the original sequence as found in the PDB and by keeping the mutations that stabilize the native structure. The native (N) and minimized (M) sequences are as follows:

TRP-cage
N: NLYIQWLKDGGPSSGRPPPS
M: PNLQTYFTLWIPSYRYPPPD
Grb2
N: TYVQALFDFDPQEDGELGFRRGDFIHVMDNSDPNWWKGACHGQTGMFPRNYVTPVNR
M: TYVQWLFQYFPKAQCYPIHIRQGFPVWVACKRKHWGIVLLQDPWCMISRNYVTNMLQ
Vav
N: GSHMPKMEVFQEYYGIPPPPGAFGPFLRLNPGDIVELTKAEAEHNWWEGRNTATNEVGWFPCNRVHPYV
M: DIDMPYCFPHHWGCAKDWMHAHRSYCSLLCHPSLAELGKWQAKGGYYWGRYSLFARDLYMQIERYHPYV.

The energy gap as well as the Z-score are estimators of the stability of the sequence in the native conformation.

Additional sequences are generated by a Monte Carlo procedure using the reference sequence as a starting point. The reference sequence s_0 is viable, since it necessarily meets the conditions listed above. At each step t, the sequence s_{t+1} is obtained by substituting one amino acid in the sequence s_t and checking that the mutated sequence still folds into the desired conformation; if it does not, s_t is reused.

When a two-class alphabet is used, the hydrophobic profile primarily determines whether a sequence does or does not fold into the target conformation (although the 3D extent of the side chain plays a role as well). The sequences generated by the Monte Carlo sampling can therefore be conveniently replaced by the hydrophobic profiles themselves. This step corresponds to a projection into the HP space. Neutral networks can be constructed by considering either the hydrophobic profiles or the more complex alphabets.

B. Infinite population model

The neutral network is an abstract object that recapitulates information regarding viable sequences and how they relate to each other. A population-dynamics model is needed to assess the likelihood for a sequence to be actually present within a population. Generally speaking, a population-dynamics model details the rules that govern births, deaths of individuals within a population, as well as the sort of mutation that can alter an individual's genome. An uneven distribution of sequence probability, as a result of the population dynamics, also entails a shift in the other properties in the population, such as average thermodynamic stability and average robustness to mutations. This shift resulting from the population dynamics will be termed "population effect." Although, in reality, the sequence space for any gene is certainly much larger than any population size, an infinite but constant population is assumed in this study, for the statistics drawn from the infinite-population model are correct as soon as $N\mu \gg 1$, where N is the population size and μ is the

mutation rate per gene and per generation.[24] While the relation $N\mu \gg 1$ does not hold in all situations, here, our focus is on the topology of the neutral network rather than on the population size, whose role was already investigated in other studies.[9,24]

We use a simple discrete-time model of evolution. In the following, $p_i(t)$ will denote the fraction of individuals which carry the sequence s_i at time t, while n_i will stand for the number of neighbors that are connected to sequence i. The quantity n_i can also be interpreted as the mutational robustness of the sequence s_i. The population at time $t+1$ is drawn from the population at time t by repeating the following operations until the size of the new generation $t+1$ has reached N.

(1) An individual is picked randomly in generation t.
(2) With probability $1-\mu$, the individual reproduces accurately and its offspring partakes in generation $t+1$.
(3) With probability μ, the individual reproduces but the gene of interest undergoes a single mutation. If the mutated sequence is viable, that is, encodes a protein that folds into the target conformation, the offspring carrying the mutated allele is included in the next generation. Otherwise, it is discarded.

One can easily derive the following expression:

$$p_i(t+1) = p_i(t) + \mu^* \left(\sum_{j \sim i} p_j(t) - \langle n \rangle \right), \tag{4}$$

where the summation involves $j \sim i$, the sequences j that neighbor the sequence i within the neutral network; μ^* is the mutation rate per generation and per amino acid; and $\langle n \rangle$ is the average mutational robustness at time t:

$$\langle n \rangle = \sum p_i(t) n_i. \tag{5}$$

Under these assumptions, regardless of the initial conditions, the population ultimately converges toward a stable steady state p^∞.[34] The vector p^∞ can be characterized mathematically as follows. If A denotes the adjacency matrix of the neutral network, p^∞ is an eigenvector associated with the highest positive eigenvalue of A, which is unique. This eigenvalue turns out to be the value $\langle n \rangle$ at steady state and will be referred to as $\langle n \rangle^\infty$. The steady state and the average mutational robustness can be efficiently computed using a shifted power method algorithm.

A convenient measure of the increase in mutational robustness is given by the "enhancement factor" ϕ:

$$\phi = \langle n \rangle^\infty / \langle n \rangle^0, \tag{6}$$

where $\langle n \rangle^0$ is the average mutational robustness when the distribution of the sequences is uniform (when $N\mu \ll 1$):

$$\langle n \rangle^0 = \frac{1}{N} \sum n_i. \tag{7}$$

For example, a value $\phi = 1.4$ means that the population dynamics increases the average mutational robustness by 40% compared to what would be expected if sequences were utilized randomly.

C. Rewiring neutral networks

Producing random graphs with a specified degree distribution is a common problem. Two main approaches exist: the matching algorithms and the switching algorithms.[35] Here, the former method is developed. By exchanging pairs of connections, it is possible to preserve a network's degree distribution while shuffling its topology. This is done by iteratively exchanging the edges between the two pairs of connected nodes, denoted (i,j) and (k,l) in this paragraph. A number of criteria must be met before the exchange is performed. (1) All nodes i, j, k, and l must be distinct. (2) The resulting network must be fully connected. Without this precaution, the network could evolve toward an ensemble of small disconnected components.

A third criterion may be added to explore topologies which favor a given property of the network. Three situations are envisaged: (3a) no additional criterion, (3b) the resulting rewired network at time $t+1$ must have a higher mutational robustness $\langle n \rangle^\infty$ at steady state than the network at time t, and (3c) the resulting rewired network at time $t+1$ must have a lower mutational robustness $\langle n \rangle^\infty$ at steady state than the network at time t. In the first case, the procedure will generate random connected networks having the desired degree distribution. In the second and third cases, the procedure will generate connected networks, biased toward topologies favoring high and low robustnesses, respectively, at steady state. Because the degree distribution is preserved, $\langle n \rangle^0$ is common to all networks generated and the enhancement ϕ will therefore evolve proportionally to $\langle n \rangle^\infty$.

III. RESULTS

The findings presented in this study are of two types: first, by using different protein models, we examine the robustness of some observations drawn from previous simulations; second, we explore the features of a neutral network that determine the enhancement of mutational robustness, ϕ, due to the population dynamics.

A. Degree distribution of the neutral network

In general, the degree distribution of a network is the first global property analyzed, for it conveys information about the composition of the network, even though it tells little about the topology itself. We first compare the degree distributions resulting from various models. Three complementary results were obtained. The first degree distributions were computed from neutral networks constructed using the 2D on-lattice protein model as described in Sec. II.[36] Figure 1(a) shows a typical degree distribution (see also Refs. 4 and 5).

The second degree distribution was computed by generating the neutral networks from the HP profiles of the sequences sampled using the 3D off-lattice protein model. For that purpose, the following simulations were conducted: for each of the proteins Vav and Grb2, a 100×10^6 step Monte Carlo simulation generated about 10^7 viable sequences (starting from the reference sequence and using the two-class energy matrix provided in Sec. II). Converting these viable

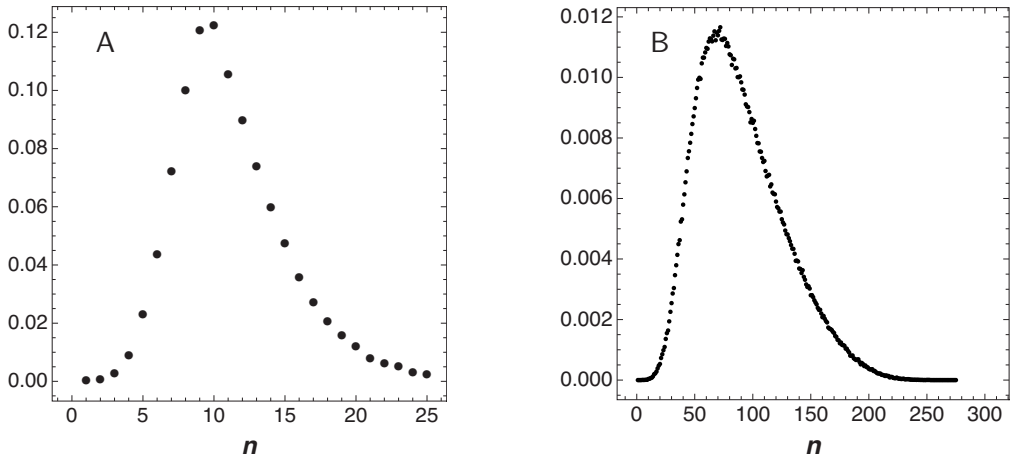

FIG. 1. (a) Degree distribution using the 2D model. The distribution shown here corresponds to the largest neutral network computed using the lattice model. (b) Degree distribution using the 3D model. The distribution shown here corresponds to the TRP-cage simulation where the number of neighbors is computed regularly along the series of sequences generated by the Monte Carlo procedure described in Sec. II. The energy matrix uses 20 classes.

sequences into HP profiles, we obtained 29 667 and 20 840 different HP profiles for Vav and Grb2, respectively. The neutral networks were computed using these profiles. The same unimodal distribution emerges from the networks, regardless of the protein considered (see Supplementary Material and Ref. 5).

It could be argued that the two previous distributions are alike because of the projection of 20-dimensional objects into a 2D space. A third distribution was therefore drawn from thorough computations carried out within the 20-class-alphabet space. About 50×10^6 sequences were generated using a 20-class energy matrix, of which 11 455 666 were viable. Every 15th sequence from this list was subjected to systematic mutation and the number of viable neighbors recorded, again using a 20-class energy matrix. By doing this, the estimation is as accurate as it can possibly be using this family of three-dimensional models. The resulting degree distribution can be seen in Fig. 1(b).

Using these two complementary models (2D and 3D), we could also establish the existence of a superfunnel. In other words, a funnel shape results when the stability of the sequences is represented as a function of their distance from the prototype sequence. This was done for the 2D protein model [see Fig. 2(a)], the HP profiles of Vav and Grb2 (see Supplementary Material[36]), and the full-alphabet model of TRP-cage. The method used in the latter case requires further explanations. The superfunnel topology is centered around the prototype sequence; though in full-alphabet space, it is probably more plausible to refer to "a group of prototype sequences." The vastness of the sequence space prevents us from directly sampling these prototype sequences. However, according to Bornberg-Bauer, the prototype sequence coincides with the consensus sequence.[18] As a result, it is reasonable to use the consensus sequence that emerged from the simulation to that purpose (NYMNNWTDGYFPRYKYPPPT). The stability of the sequences is measured, thanks to the *Z*-score as defined in Sec. II. The results are presented in Fig. 2(b).

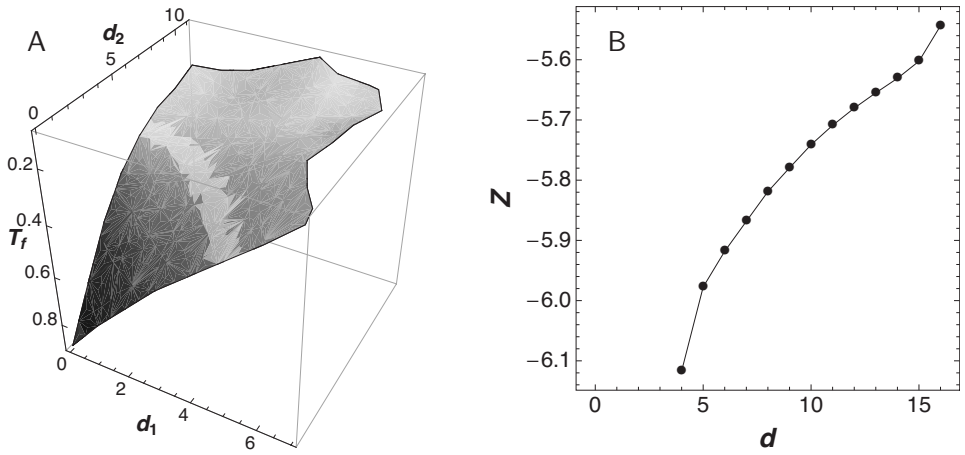

FIG. 2. (a) Superfunnel using the 2D model. The plot represents the stability (as measured by the folding temperature T_f) as a function of two coordinates $\{d_1,d_2\}$ that express the distance from the prototype sequence. The neutral network used is the largest network obtained with the on-lattice model (67 614 sequences). For a given sequence, d_1 is the number of hydrophobic residues that are polar in the prototype sequence, while d_2 is the number of polar residues that are hydrophobic in the prototype sequence. The z axis is oriented downward to conform to the classical "funnel" picture. (b) Superfunnel using the 3D model. The plot represents the stability of a sequence as a function of the Hamming distance from the consensus sequence NYMNNWTDGYFPRYKYPPPT (see text for details). The stability is measured by the *Z*-score of the native conformation with respect to the decoy structures.

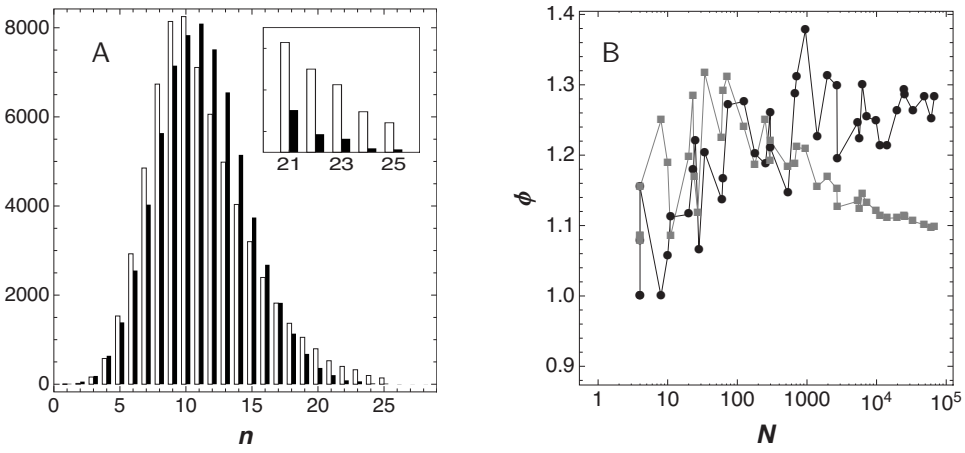

FIG. 3. (a) Degree distribution of two matching networks. The original neutral network was drawn from the 2D protein model. It is composed of 67 614 sequences connected through 378 673 edges. Its degree distribution is represented by the white bars. A random network of the same size was prepared using the Erdős–Rényi random graph model (equal number of nodes and edges). Its degree distribution is represented by the black bars. Inset: zoom of the tail of the distribution. (b) Mutational robustness's enhancement as a function of the network size N. In black, ϕ as a function of N for neutral networks drawn from the 2D protein model. In gray, the same curve for equally large random networks built according to the Erdős–Rényi model.

B. Topological features favoring mutational robustness

1. Neutral networks versus random networks

"Random networks" may be prepared in many ways depending on the properties that are desired. The study of random networks originates from the late 1950s, with Erdős and Rényi's model.[37] Even though this model is rarely used in practice, for most real-life networks significantly depart from it, it is still useful to set a null model for comparison to our neutral networks.

Using the 2D model, 2977 neutral networks were constructed and the enhancement ϕ of mutational robustness through population dynamics computed for 40 of them (sizes range from 1 to 67 614). In parallel, 40 random networks were constructed according to Erdős and Rényi's model. This was done in order for their size and connection number to match roughly the neutral networks'. Figure 3(a) presents an example of degree distribution of two matching networks. The two distributions look fairly similar, apart from the presence of rare but essential "hubs"—highly connected nodes—in the neutral network [see inset of Fig. 3(a)].

The enhancements of the mutational robustness through population dynamics for these random networks were subsequently computed. Figure 3(b) shows how the enhancement ϕ of Erdős–Rényi random graphs compares to the neutral networks' as a function of the network size.

Figure 3(b) suggests that, as size increases, a neutral network departs more and more from the Erdős–Rényi model, with respect to the enhancement ϕ. Up to a size $N=1000$, the enhancements in both situations remain fairly comparable and increase as size increases. However, beyond $N=1000$, while it still increases for neutral networks, the enhancement computed on Erdős–Rényi graphs starts to drop down.

The existence of hubs seems to be of importance as well: similar computations were carried out for scale-free networks generated using the algorithm proposed by Babarási and Albert.[38] For equivalent sizes, scale-free networks can lead to much higher enhancements ϕ, up to 10 within the size range considered, than neutral networks (see Supplementary Material[36]). However, the number of edges in a scale-free network of size N is much lower than a neutral network's: the former is of the order of N whereas the latter increases as a function of $N \log N$ (data not shown). This observation considerably limits the scope our comparison between scale-free and neutral networks.

2. Autocorrelation of the mutational robustness

The results discussed in Sec. III B 1 naturally prompt us to investigate the topological features underlying high enhancements of mutational robustness. Since the degree distribution seems remarkably stable across the different protein models, we generated random networks that had a degree distribution typical of neutral networks. This was achieved by using a switching algorithm,[35] as described in Sec. II. This approach generates random networks with a given degree distribution by rewiring randomly the network. The original network, at time $t=0$, is a 504-node network obtained with the 2D model. Two simulations were performed that allowed us to direct the rewired networks toward high-enhancement and low-enhancement topologies. Another simulation was left unconstrained. In each case, 500 000 Monte Carlo steps were performed.

The evolution of the mutational robustness $\langle n \rangle^{\infty}$ as a function of the number of Monte Carlo steps is shown in Fig. 4(a). The unconstrained simulation as well as the directed downward simulation have converged after 500 000 steps, and the directed upward one is close enough to convergence. The values $\langle n \rangle^{\infty}$ attained range from 6.520 to 10.713. The typical values sampled when rewiring unconstrainedly the network are around 6.594 ± 0.025 (standard deviation). Interestingly, the mutational robustness computed for the original neutral network is 6.924, which is significantly higher than that of the networks generated by the unconstrained Monte Carlo procedure. A neutral network's topology is therefore prone to more marked effects, owing to the population dynamics and to the selection of sequences robust to mutations.

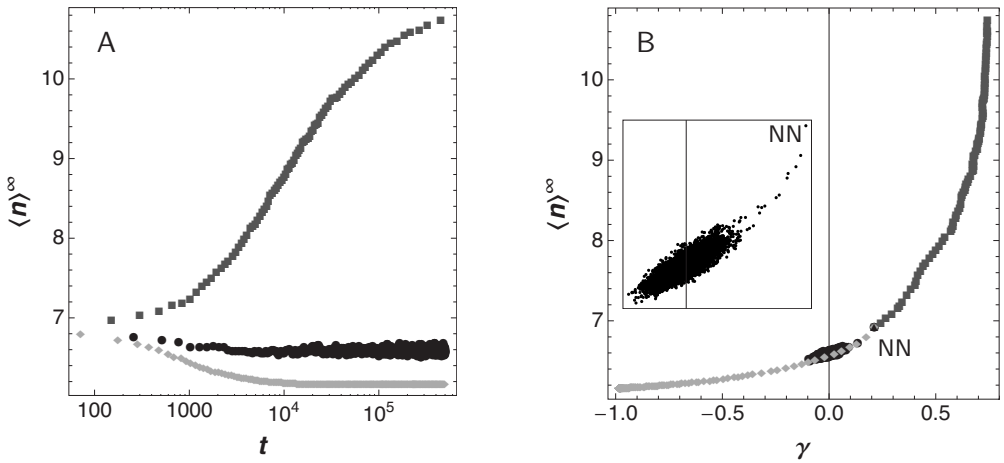

FIG. 4. (a) Evolution of the mutational robustness at steady state $\langle n\rangle^{\infty}$ as a function of the number t of edges exchanged. Three simulations are performed: (a) the edges are exchanged without any other constraint than the network's full connectivity (in black); (b) the edges are exchanged only if the mutational robustness in the resulting network increases (in dark gray); (c) the edges are exchanged only if the mutational robustness in the resulting network decreases (in light gray). (b) Mutational robustness at steady state $\langle n\rangle^{\infty}$ as a function of the local autocorrelation γ within a network. During the three simulations discussed above, networks were collected periodically to perform further analyses (unconstrained in black, increasing robustness in dark gray, and decreasing robustness in light gray; the original neutral network is represented by a black triangle, labeled "NN"). In particular, we can compute the mutational robustness at steady state, $\langle n\rangle^{\infty}$ as a function of the autocorrelation γ. Because ϕ is proportional to $\langle n\rangle^{\infty}$, any property of $\langle n\rangle^{\infty}$ can be extended to ϕ.

To clarify the main topological feature underlying these higher population effects, we considered the correlation that exists between the mutational robustnesses of neighbors within the network. This correlation, hereafter denoted γ, is defined by

$$\gamma = \frac{\langle (n_i - \langle n\rangle)\cdot(\langle n_j\rangle_{j\sim i} - \langle n\rangle)\rangle_i}{\mathrm{var}(n)}, \tag{8}$$

where $\langle\cdots\rangle_k$ symbolizes the average over the indices k and $j\sim i$ indicates the neighbors of sequence i within the network. The correlation γ indicates how smoothly the mutational robustness varies across the neutral network. A value close to 1 means that a highly connected node is generally surrounded by highly connected nodes and, conversely, a poorly connected node is generally surrounded by poorly connected nodes (the mutational robustness varies smoothly). A value close to −1 indicates that highly connected nodes are surrounded by poorly connected nodes and vice versa (the mutational robustness varies extremely ruggedly but regularly). Finally, a value close to 0 indicates the absence of any clear influence of one node on its neighbors (the mutation robustness varies ruggedly).

The results for all three simulations are represented in Fig. 4(b). The correlation γ varies from −1 to +0.8. There appears to be a simple relation between γ and ϕ. The unconstrained simulation leads to correlation values γ, located around zero, indicating that this simulation can effectively maintain a given degree distribution and lose the correlation that was initially present in the neutral network. The directed simulations clearly demonstrate that the correlation across the network, with respect to the mutational tolerance, is the main factor in determining the extent of the enhancement factor ϕ. Interestingly, all the results drawn from the three simulations perfectly line up. Remarkably, this is also the case of the unconstrained simulation (see the inset in Fig. 4(b)). The relation is thus unlikely to result from the artificial and stringent selection operated during the directed simulations.

IV. CONCLUSIONS AND DISCUSSION

Since Kimura's seminal work,[17] it has become common to assume that the vast majority of mutations that are not lethal confer little or no functional advantage. Yet, the picture of a completely random genetic drift on a neutral network had to be altered: when the mutation rate becomes high, sequences that are robust to mutations are positively selected. Apart from the mutation rate μ, the other elements that determine the intensity of the selection for higher mutational robustness are the population size N and the organization of the viable sequences in sequence space. By favoring mutationally robust sequences, selection increases the average mutational robustness and accordingly reduces the number of unproductive lethal mutations (mutational load).

On the one hand, the viable sequences were shown to be organized in a superfunnel;[18,19] on the other hand, the influence of the mutation rate μ, the population size N, or the sorts of mutations considered on the strength of the selection for higher mutational robustness was investigated.[4,9,24] Nonetheless, the salient topological features of the superfunnel that could affect the strength of the selection remained to be explored. The question of these features was addressed in this study.

Assuming that functionality coincides with thermodynamic stability, two different models led to a clear consensus picture of what the typical degree distribution of a neutral network must look like. The distribution consistently appears unimodal. According to these results, a sequence is rarely very poorly connected. Also, the superfunnel topology emerges from all models: from the on-lattice model and the off-lattice model, considering either HP profiles or full-

alphabet sequences.

One distinctive feature of the superfunnel topology is the existence of sequences that are highly robust to mutations. These sequences, which have many neighbors within the neutral networks, are commonly referred to as hubs. Random networks generated using the model of Erdős and Rényi do not possess hubs. We showed that the reduction of the mutational load is more marked in the case of the neutral networks, as soon as the size of the networks is large enough. In contrast, hubs are frequent in scale-free random networks generated using Barabási and Albert's method.[38] In these networks, the selection for mutational robustness is extremely marked. Hubs therefore appear to be involved in increasing the selection for mutational robustness and in the subsequent decrease in mutational load. However, whereas the Erdős–Rényi random networks were prepared so as to have a comparable number of connections, scale-free networks are much sparser networks. Their limited number of connections may artificially limit the sequences' ability to drift away from the hubs.

Within the superfunnel, it has to be expected that the neighbors of a mutationally robust sequence are robust too and that, conversely, the neighbors of a mutationally nonrobust sequence are nonrobust. Mathematically, the mutational robustness of a sequence is correlated with that of the sequence's neighbors. With respect to the selection of mutationally robust sequences, the relevance of the correlation γ in mutational robustness was further investigated by generating random networks having the same degree distribution as a neutral network. The networks generated, in particular, possess the same number of hubs as the original neutral network. We related the factor γ to the mutational robustness $\langle n \rangle^{\infty}$ and demonstrated that the network's "smoothness" γ directly affects the mutational robustness $\langle n \rangle^{\infty}$ [see Fig. 4(b)]. Since ϕ and $\langle n \rangle^{\infty}$ are directly proportional, the network's smoothness γ directly affects the increase in mutational robustness ϕ too.

In relating the correlation factor γ and the mutational robustness, the networks we used were rather small because of computational limitations. It would be valuable to extend our comparison between the neutral and random networks to larger network sizes. It also has to be remarked that Erdős–Rényi random networks also differ from neutral networks by the lack of smoothness. The difference observed between neutral networks and Erdős–Rényi random networks is a combined effect of the absence of hubs and smoothness in the latter.

The correlation between ϕ and γ is striking. The relation obtained through computations is very clear, even in the case of the unconstrained simulation, and it demonstrates that one can relate a global characteristic of neutral networks ϕ, which is an eigenvalue of the graph, to a local characteristic γ, which tells how likely is a sequence to be as robust as its neighbors. This observation is important for potential applications of this theory: the mutational robustness enhancement can hardly be experimentally estimated, for one should cover most of the neutral network of a gene. However, it might be assessed through the correlation γ observed for a limited but statistically large enough number of sequences (experimentally or computationally using software such as FOLDX[39]). Alongside other factors,[10] this may be useful, for instance, when designing structures *de novo*.

ACKNOWLEDGMENTS

The authors express their gratitude to N. Lartillot (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier) for fruitful discussions.

[1] Y. Xia and M. Levitt, Proteins **55**, 107 (2004).
[2] H. S. Chan and E. Bornberg-Bauer, Appl. Bioinf. **1**, 121 (2002).
[3] J. L. England, B. E. Shakhnovich, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **100**, 8727 (2003).
[4] Y. Xia and M. Levitt, Proc. Natl. Acad. Sci. U.S.A. **99**, 10382 (2002).
[5] J. Noirel and T. Simonson, BMC Bioinf. **7**, 79 (2007).
[6] K. B. Zeldovich, P. Chen, B. E. Shakhnovich, and E. I. Shakhnovich, PLOS Comput. Biol. **3**, e139 (2007).
[7] C. O. Wilke, BMC Genet. **5**, 25 (2004).
[8] W. Fontana and P. Schuster, Science **280**, 1451 (1998).
[9] R. Forster, C. Adami, and C. O. Wilke, J. Theor. Biol. **243**, 181 (2006).
[10] J. D. Bloom, C. O. Wilke, F. H. Arnold, and C. Adami, Biophys. J. **86**, 2758 (2004).
[11] B. P. Blackburne and J. D. Hirst, J. Chem. Phys. **115**, 1935 (2001).
[12] B. P. Blackburne and J. D. Hirst, J. Chem. Phys. **123**, 154907 (2005).
[13] J. Maynard Smith, Nature (London) **225**, 563 (1970).
[14] D. J. Lipman and W. J. Wilbur, Proc. R. Soc. London **245**, 7 (1991).
[15] F. L. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[16] F. L. Lau and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **87**, 638 (1990).
[17] M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1983), reprinted in 1986.
[18] E. Bornberg-Bauer, Biophys. J. **73**, 2393 (1997).
[19] E. Bornberg-Bauer and H. S. Chan, Proc. Natl. Acad. Sci. U.S.A. **96**, 10689 (1999).
[20] Y. Xia and M. Levitt, Curr. Opin. Struct. Biol. **14**, 202 (2004).
[21] R. Wroe, E. Bornberg-Bauer, and H. Chan, Biophys. J. **88**, 118 (2005).
[22] D. M. Taverna and R. A. Goldstein, Proteins **46**, 105 (2002).
[23] S. Bershtein, K. Goldin, and D. S. Tawfik, J. Mol. Biol. **379**, 1029 (2008).
[24] E. van Nimwegen, J. P. Crutchfield, and M. Huynen, Proc. Natl. Acad. Sci. U.S.A. **96**, 9716 (1999).
[25] D. M. Taverna and R. A. Goldstein, J. Mol. Biol. **315**, 479 (2002).
[26] J. Noirel, "Évolution in silico des protéines monomériques et dimériques," Ph.D. thesis, École Polytechnique, 2005.
[27] H. Li, C. Tang, and N. S. Wingreen, Proteins **49**, 403 (2002).
[28] G. Launay, R. Mendez, S. Wodak, and T. Simonson, BMC Bioinf. **8**, 270 (2007).
[29] L. Wernisch, S. Hery, and S. J. Wodak, J. Mol. Biol. **301**, 713 (2000).
[30] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[31] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[32] H. S. Chan, Nat. Struct. Biol. **6**, 994 (1999).
[33] P. Tuffery, C. Etchebest, S. Hazout, and R. Lavery, J. Biomol. Struct. Dyn. **8**, 1267 (1991).
[34] M. Eigen, Naturwiss. **58**, 465 (1971).
[35] R. Milo, N. Kashtan, S. Itzkovitz, M. Newman, and U. Alon, e-print arXiv:cond-mat/0312028.
[36] See EPAPS Document No. E-JCPSA6-129-006839 for the energy matrices, the degree distributions for the 3D model, and the results about scale-free networks. For more information on EPAPS, see http://www.aip.org/pubservs/epaps.html.
[37] P. Erdős and A. Rényi, Publ. Math. (Debrecen) **6**, 290 (1959).
[38] A. L. Barabási and R. Albert, Science **286**, 509 (1999).
[39] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano, The FoldX web server: an online force field (2005) [Nucleic Acids Res. **33**, W382 (2005)].