Evolution of sparsity and modularity in a model of protein allostery

Mathieu Hemery[1,2,3] and Olivier Rivoire[2,3,*]
[1]*ESPCI ParisTech, PCT, Gulliver, F-75005, Paris, France*
[2]*CNRS, LIPhy, F-38000 Grenoble, France*
[3]*Univ. Grenoble Alpes, LIPhy, F-38000 Grenoble, France*
(Received 1 October 2014; published 9 April 2015)

The sequence of a protein is not only constrained by its physical and biochemical properties under current selection, but also by features of its past evolutionary history. Understanding the extent and the form that these evolutionary constraints may take is important to interpret the information in protein sequences. To study this problem, we introduce a simple but physical model of protein evolution where selection targets allostery, the functional coupling of distal sites on protein surfaces. This model shows how the geometrical organization of couplings between amino acids within a protein structure can depend crucially on its evolutionary history. In particular, two scenarios are found to generate a spatial concentration of functional constraints: high mutation rates and fluctuating selective pressures. This second scenario offers a plausible explanation for the high tolerance of natural proteins to mutations and for the spatial organization of their least tolerant amino acids, as revealed by sequence analysis and mutagenesis experiments. It also implies a faculty to adapt to new selective pressures that is consistent with observations. The model illustrates how several independent functional modules may emerge within the same protein structure, depending on the nature of past environmental fluctuations. Our model thus relates the evolutionary history of proteins to the geometry of their functional constraints, with implications for decoding and engineering protein sequences.

DOI: 10.1103/PhysRevE.91.042704 PACS number(s): 87.14.E−, 87.23.Kg, 05.40.−a

I. INTRODUCTION

Natural proteins are well known to be highly tolerant to point mutations: most of the amino acids forming their sequence can be changed without affecting notably their biochemical properties [1]. Statistical analysis of protein sequences, which study mutational patterns over large data sets of natural protein sequences [2], and saturated mutagenesis experiments, which assay every single mutation on a particular protein [3], are now revealing the spatial architecture of this robustness: in several proteins, the amino acids most essential to the function have been found to be organized in small, structurally connected clusters of interacting and coevolving residues, called protein sectors [4].

For instance, in PDZ domains, a family of small interaction domains that are often part of larger protein complexes, a sector connects the ligand binding pocket to an opposite surface site [2,3]. An interaction with another protein at the same surface site has been shown to control the affinity of the ligand at the binding pocket [5], leading to the hypothesis that the sector serves to transmit information between the two sites. Such a regulation of one site on the protein by another, distant site is called "allostery" [6]. Similar sectors have been found and experimentally investigated in other proteins, where they also consist of small structurally connected subsets of residues and, in many cases, mediate allostery [7–9]. Furthermore, several quasi-independent sectors have been reported to coexist within the same protein domain [4].

The generality of the concept of protein sector will require further work to be generally established, but the very heterogeneous distribution of functional constraints within a protein structure is indisputable. For instance, residues in the interior of protein structures systematically tend to be evolutionarily more conserved. As these residues are in contact with more other residues than those closer to the surface, this observation may reflect a larger number of physical constraints. More generally, the heterogeneous distribution of functional constraints may be inherent to the physical properties of proteins, including the functional properties for which they were selected. For instance, when the function involves binding to a ligand, the residues structurally closer to the ligand may be expected to be functionally more important. We shall show, however, that such structural heterogeneities are not needed to explain a spatial concentration of functional constraints within a protein structure. To this end, we introduce below a simple mathematical model in which all "residues" are *a priori* equivalent, but where a sparse sector can nevertheless arise as a consequence of fluctuations during the evolutionary process.

The role of evolutionary history in shaping biological organizations has been discussed previously. In particular, it has been studied in relation to "modularity," the generic decomposition of biological networks into subnetworks [10]. Explanations for the evolutionary origin of modularity broadly fall in two classes [11]: first, those based on the combinatorial properties of the process generating new variations, e.g., gene duplications and recombinations [12], and, second, those invoking the history of selective pressures, notably the particular structure of environmental fluctuations [13].

Proteins are modular at several levels. First, they may be composed of several domains, which are subparts of a protein that can fold independently into stable units [14]. The sequence of a multidomain protein thus consists of the concatenation of the sequences of different domains. The duplication, divergence, and recombination of these domains is a major source of new proteins in evolution [15]. Second, a given domain may itself be composed of one or several

*olivier.rivoire@ujf-grenoble.fr

submodules called sectors [4]. These modules are structurally connected, but, at variance with protein domains, the amino acids of distinct sectors are entangled along the sequence so that they cannot easily be combinatorially reassorted by recombination; instead, selective pressures may play a more dominant role in their formation and evolution.

In general, however, the variational and selective factors are nonexclusive and may contribute jointly to the emergence of modules [16]. Beyond the question of their origin, the implications of modular architectures for future evolution have also been extensively studied in terms of resilience to mutations, or "robustness," and in terms of faculty to adapt, or "evolvability" [17], two properties found to have a complex relationship [18–20].

The presence of a single sector in the network of interacting residues forming the structure of a protein, which typically comprises only $\sim$20% of the total number of amino acids, corresponds to a degenerate form of modularity, better referred to as "sparsity" [21]. Very generally, a network is said to be sparse when only a small fraction of the possible links between pairs of nodes is present; in the present context where we consider the network of functional interactions between residues, sparsity refers to the fact that only a small fraction of the specific interactions between amino acids is essential for the function of a protein.

We demonstrate, in the context of a physical model of protein evolution, how sparsity generically emerges in the form of a spatial concentration of functional constraints from fluctuations during the evolutionary process. These fluctuations may involve variational or selective factors and may promote robustness and/or evolvability to varying degrees. The phenomenon that we describe is more elementary than the evolution of modularity, which arises when the fluctuations have some additional structure, in relation to the structure of the function itself.

II. A MODEL FOR THE EVOLUTION OF ALLOSTERY

A. Physical model

To illustrate the role of evolutionary history in a context where structural heterogeneities are minimized, we introduce a model defined on a regular structure and consider an allosteric property, which may in principle involve the entire structure. In the spirit of previous theoretical studies of protein evolution [22], we present this model in the generic framework of spin glasses [23], but the Gaussian spin glass [24] that we analyze more specifically is also closely related to models of elastic networks commonly used to study protein dynamics [25].

We may derive our model starting from a general expression for the energy of a protein,

$$E = -\sum_i K_0(a_i,\sigma_i,\varepsilon(r_i)) - \sum_i K_1(a_i,a_{i+1},\sigma_i,\sigma_{i+1}) - \sum_{i,j} K_2(a_i,a_j,r_i,r_j,\sigma_i,\sigma_j), \quad (1)$$

where a_i indicates the amino acid at position i along the chain, r_i its mean position, e.g., of its alpha carbon, and σ_i its physical state, e.g., its orientation and fluctuations around the mean position. The term $K_0(a_i,\sigma_i,\varepsilon(r_i))$ represents an interaction energy between the amino acid a_i and its local environment $\varepsilon(r_i)$, which may include the solvent and/or the amino acids of another protein, $K_1(a_i,a_{i+1},\sigma_i,\sigma_{i+1})$ represents the bonding energy between successive amino acids along the chain, and $K_2(a_i,a_j,r_i,r_j,\sigma_i,\sigma_j)$ the interactions of residues far apart along the chain but brought together upon folding.

While others have studied the incidence of evolutionary parameters on protein structure and stability [26,27], we focus here on the evolution of amino-acid specific variables in the context of a fixed structure, to analyze the structural organization of functional constraints in members of a protein family sharing a common fold. We thus fix the r_i at the nodes of a lattice, where only nearest neighbors have nonzero interactions. For simplicity, and to minimize structural heterogeneities, we ignore in this context the distinction between bond and nonbond energies, so that

$$E = -\sum_{\langle i,j\rangle} K(a_i,a_j,\sigma_i,\sigma_j) - \sum_i K_0(a_i,\sigma_i,\varepsilon(r_i)), \quad (2)$$

where $\langle i,j\rangle$ indicates neighboring sites on the lattice (neighborhood relationships thus define the range of the interactions).

We further assume that the σ_i take real values and that K has the form $K(a_i,a_j,\sigma_i,\sigma_j) = J(a_i,a_j)\sigma_i\sigma_j$. Similarly, we assume that the environment around i is represented by a real number h_i with K_0 of the form $K_0(a_i,\sigma_i,\varepsilon(r_i)) = h_i\sigma_i$ (more generally, h_i could depend on a_i). We thus arrive at an energy of the form

$$E(\sigma|J,h) = -\sum_{\langle i,j\rangle} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i, \quad (3)$$

which is formally the energy of a spin glass [28], where spins σ_i interact in the context of given (quenched) couplings $J_{ij} = J(a_i,a_j)$ and fields h_i.

More specifically, we shall consider a Gaussian spin glass model, defined at inverse temperature β by the partition function [24]

$$Z(J,h) = \int \prod_i \frac{e^{-\sigma_i^2/2}}{\sqrt{2\pi}} d\sigma_i \; e^{-\beta E(\sigma|J,h)}, \quad (4)$$

where the energy $E(\sigma|J,h)$ given by Eq. (3) can also be written as $E(\sigma|J,h) = -\frac{1}{2}\sigma^\top J\sigma - h^\top\sigma$, with σ representing a vector $\sigma = (\sigma_1\ldots,\sigma_M)$ and with the geometry of the lattice defined by the nonzero elements of the matrix J_{ij} (with $J_{ii} = 0$ and $J_{ij} = J_{ji}$). This model is defined only at high temperature since the integral diverges for large β, but if c denotes the maximal connectivity of the lattice, it is sufficient to assume that $|\beta h_i|, |\beta J_{ij}| < 1/c$ for the integral to converge.

The simplifications leading to the Gaussian spin glass model are drastic but retain the essential relationships between the variables of the problem: the couplings J_{ij}, which may be subject to evolution, the environmental variables h_i, which may vary with time, and the physical variables σ_i, which are subject to short-range interactions constrained by the overall structure and are dependent on the identity of the amino acids and on the environment. The model may also be viewed as an elastic network model [25] with a single degree of freedom per

site and nonuniform "spring constants." Remarkably, despite their extreme simplicity, elastic network models are known to capture some of the functional properties of natural proteins [29].

The Gaussian model is analytically solvable, with

$$Z(J,h) = [\det(I - \beta J)]^{-1/2} \exp\left[\frac{\beta^2}{2} h^\top (I - \beta J)^{-1} h\right], \tag{5}$$

where I represents the $M \times M$ identity matrix, M being the number of nodes in the lattice. Its free energy $F(J,h) = -\beta^{-1} \ln Z(J,h)$ therefore has the form

$$F(J,h) = -\tfrac{1}{2}\beta\, h^\top (I - \beta J)^{-1} h + F(J,0), \tag{6}$$

where $F(J,0)$ does not depend on h. This expression allows us to calculate a free energy of binding with an external ligand: the presence of a ligand corresponds indeed to a field h' differing from the field h in its absence, so a binding free energy is obtained as the difference $F(J,h') - F(J,h)$. On the other hand, we can also calculate the difference of free energy due to mutations, whose effect is to change J into J', as $F(J',h) - F(J,h)$.

Here we consider a cylindrical lattice where two different ligands can bind at the two opposite open ends [Fig. 1(a)]. The cylindrical lattice is chosen because its regular shape represents the least favorable geometry for the emergence of structural heterogeneities. We quantify the preferential binding of one of the ligands in the presence of the other, which corresponds to allosteric regulation [6]. Following the terminology used for allosteric proteins, we call "regulatory site" (abbreviated in "reg") the upper end of the cylinder, "modulator" the ligand binding to it, "active site" ("act") its lower end, and "endogenous ligand" the ligand binding to it. Taking the interaction of site i with the solvent to correspond to $h_i = 0$, we thus have

$$E(\sigma|J,h) = -\sum_{\langle i,j\rangle} J_{ij}\sigma_i\sigma_j - \sum_{i\in\text{reg}} h_i^{\text{reg}}\sigma_i - \sum_{i\in\text{act}} h_i^{\text{act}}\sigma_i, \tag{7}$$

where $h_i^{\text{reg}} = m_i$ in the presence of a modulator characterized by the vector m_i ($i \in$ reg), $h_i^{\text{reg}} = 0$ in its absence, and $h_i^{\text{act}} = \ell_i$ in the presence of an endogenous ligand characterized by the vector ℓ_i ($i \in$ act), $h_i^{\text{act}} = 0$ in its absence.

Allostery corresponds to a more favorable interaction with a ligand ℓ in the presence of a modulator m. It is quantified thermodynamically in terms of free energy differences [30], as

$$\phi(J,\ell,m) = \Delta F(J,\ell|0) - \Delta F(J,\ell|m), \tag{8}$$

where $\Delta F(J,\ell|0) \equiv F(J,h^{\text{act}} = \ell, h^{\text{reg}} = 0) - F(J,h^{\text{act}} = 0, h^{\text{reg}} = 0)$ represents the binding free energy of ℓ in absence of m, and $\Delta F(J,\ell|m) \equiv F(J,h^{\text{act}} = \ell, h^{\text{reg}} = m) - F(J,h^{\text{act}} = 0, h^{\text{reg}} = m)$ in its presence, as illustrated in Fig. 1(a) (these definitions are for a given J, representing a given sequence of amino acids). In the context of our Gaussian model, using Eq. (6) leads to an explicit expression for allosteric efficiency,

$$\phi(J,\ell,m) = \beta\, m^\top [(I - \beta J)^{-1}]_{\text{reg,act}}\, \ell, \tag{9}$$

where $[A]_{\text{reg,act}}$ denotes a submatrix of A_{ij} where i is restricted to $i \in$ reg and j to $j \in$ act.

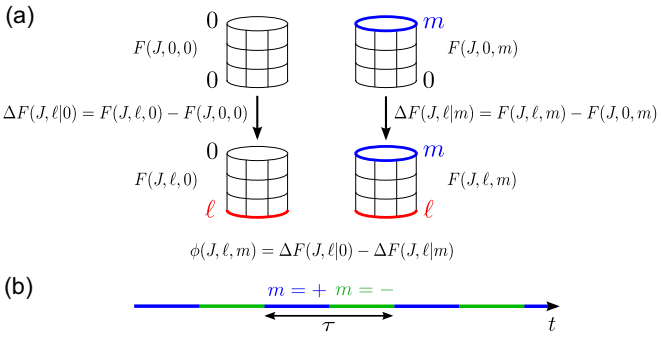


FIG. 1. (Color online) (a) A model of protein allostery is defined as a spin glass on a cylindrical lattice, a geometry chosen because of its structural homogeneity. In this model, spins σ_i are on the nodes to represent physical variables, and couplings J_{ij} are on the edges to represent interactions between these variables. These J_{ij} are subject to evolution. Interactions with a modulator m and/or a ligand ℓ are modeled by fields h_i applied to the sites i at the open ends of the cylinder. Allostery is quantified by $\phi(J,\ell,m)$, the difference between the binding free energy of ℓ in the presence of m, $\Delta F(J,\ell|m)$, and in its absence, $\Delta F(J,\ell|0)$. (b) The sequences of the modulator and ligand are defined by the values and signs of the fields h_i, with $h_i = 0$ representing an interaction with the solvent. Evolution is performed over a population of systems with selection for allosteric efficiency and with mutations affecting the couplings J_{ij} at a rate μ. A given generation is selected for allostery with given ℓ,m, but when the environment fluctuates, different generations may experience different ℓ,m. By symmetry, varying ℓ or m is equivalent, and we fix the sequence of the ligand to $h_i = +1$ for all i at the bottom of the cylinder when ℓ is present and vary only the sequence of the modulator every $\tau/2$ generations, between $h_i = +1$ for i at the top ($m = +$) and $h_i = -1$ ($m = -$). Modulators or ligands with nonuniform sequences also can be considered as in Fig. 9.

B. Evolutionary dynamics

To study how allostery may be implemented through evolution, we perform numerical simulations of an evolutionary dynamics using a standard genetic algorithm [31]. We start with a large population of $P = 500$ systems, with random couplings defined on 10×10 square lattices, and repeat cycles of selection, reproduction, and mutation for a large number (5×10^4) of generations. Selection and reproduction are based on allosteric efficiency, as defined by Eq. (8), with systems with larger allosteric efficiency generating more offspring. Specifically, a system k with couplings $J_{ij}^{(k)}$ is replicated n_k times based on the value of $\phi_k = \phi(J^{(k)},\ell,m)$, following the sigma-scaling rule [31], $n_k = 1 + (\phi_k - \overline{\phi})/(2\sigma_\phi^2)$, where $\overline{\phi}$ and σ_ϕ^2 are, respectively, the mean and variance of ϕ in the population. This particular relation between n_k and ϕ_k is convenient, but its particular form is not determining: what is essential is that larger values of ϕ_k imply larger values of n_k; we verified, for instance, that an elite strategy, which is another standard rule used with genetic algorithms [31], has equivalent implications.

Mutations of the amino acids correspond to changes of the couplings; for simplicity, instead of introducing an arbitrary matrix $J(a,b)$, we assume that a mutation randomly changes the value of a single coupling $J_{ij} = J(a_i,a_j)$ at a rate μ per generation, independently of the other couplings: more precisely, each coupling J_{ij} has a probability μ to be mutated to a random value in $[-1,1]$ [in what follows, we fix the inverse temperature to $\beta = 0.1$ so that the integral in Eq. (4) is always well defined]. We verified, however, that explicitly mutating amino acids at the level of sites, which affect simultaneously several couplings, led to similar results.

Numerical simulations of evolutionary dynamics are generally limited by the computational cost of estimating the fitness of each individual. The Gaussian model, which is analytically solvable, is thus particularly well suited to an evolutionary analysis: it allows us to perform efficient calculations in the context of an arbitrary geometry.

III. EVOLUTIONARY CONCENTRATION OF FUNCTIONAL CONSTRAINTS

A. Sparsity

The outcome of the evolutionary dynamics is contingent on the series of modulator and ligand sequences that the successive generations encounter [Fig. 1(b)]. When these sequences are constant over time, say, $m = (+1,\ldots,+1)$ and $\ell = (+1,\ldots,+1)$ at all times, systems evolve maximal couplings $|J_{ij}| \simeq 1$ at all sites. This implementation of the couplings optimizes the allosteric efficiency ϕ and epitomizes an absence of sparsity. Repeating the same simulations with a modulator that alternates with period τ between two sequences, $m = (+1,\ldots,+1)$ and $m = (-1,\ldots,-1)$, yields a qualitatively different outcome: the smaller τ is, the fewer are the large couplings, as illustrated in Fig. 2(a).

Allostery requires strong couplings, but not all strong couplings need to be functionally significant: if a strong coupling is defined by $|J_{ij}| > 0.8$ as in Fig. 2(a), we may expect $\sim$20% of strong couplings even in absence of any selection, only because the J_{ij} are mutated to random values in $[-1,1]$ (0.8 is an arbitrary cutoff but other values lead to a similar conclusion; see Fig. 3).

As a more relevant measure of functional significance, we may consider instead the "fitness cost" $\delta\phi_{ij}$ that a mutation of J_{ij} can cause to the allosteric efficiency ϕ. To compare systems with different allosteric efficiencies, we define here the relative fitness cost of a mutation $J_{ij} \to J^*_{ij}$ as $\delta\phi^*_{ij}(J) \equiv [\phi(J) - \phi(J^{(*)})]/\phi(J)$, where $J^{(*)}$ differs from J by the value of J_{ij}. Figure 2(b) shows the result of retaining only the couplings with largest effect when mutated, with $\delta\phi_{ij} > 0.1$ (other choices of this cutoff yield similar results; see Fig. 3). This criterion, closer to what has been experimentally measured [3], reveals distinctly the presence of a connected subset of functional couplings joining the regulatory and active sites.

The subsets of functionally significant couplings shown in Fig. 2(b), which break the rotational invariance of the cylinder and whose locations vary from simulation to simulation, display several features reminiscent of protein sectors observed in natural proteins [2,4]: (i) they are overall structurally connected [Fig. 2(b)]; (ii) they have a hierarchical organization:

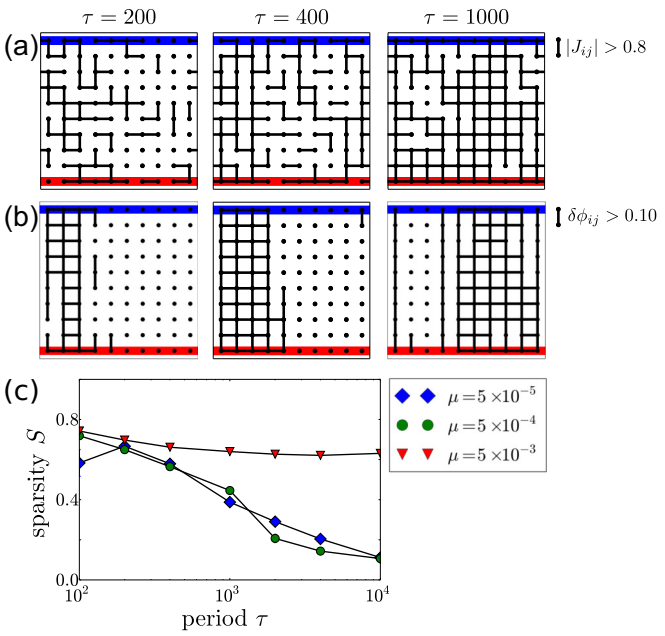


FIG. 2. (Color online) Examples of systems obtained from an evolutionary dynamics with mutation rate $\mu = 5 \times 10^{-5}$ and different periods τ of fluctuations of selective pressure ($\tau = 200,400,1000$). (a) Couplings J_{ij} with large absolute values, $|J_{ij}| > 0.8$. (b) Couplings J_{ij} inducing a large loss in allosteric efficiency when mutated, $\delta\phi_{ij} > 0.1$ (see Fig. 3 for other values of the cutoffs). The figures display the fittest individual in a population of $P = 500$ individuals prior to a change of environment. (c) Sparsity of evolved systems as a function of the period τ of environmental changes, where sparsity is defined as the fraction of nonrepresented couplings in (b). The error-bars (standard deviations over 10 simulations) are smaller than the marker.

less significant couplings are peripheral to more significant ones, as shown by varying the value of the cutoff defining functional significance (Fig. 3); (iii) they are evolutionarily conserved: their location is stable over multiple periods along a given evolutionary trajectory (Fig. 4); (iv) their couplings

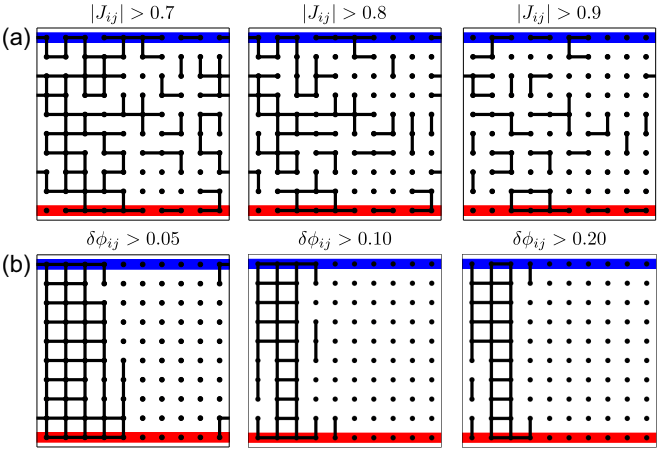


FIG. 3. (Color online) For the system associated with $\tau = 200$ in Fig. 2, couplings $|J_{ij}|$ and functional constraints $\delta\phi_{ij}$ above different values of the cutoffs (Fig. 2 corresponds to the middle column, $|J_{ij}| > 0.8$ and $\delta\phi_{ij} > 0.1$).

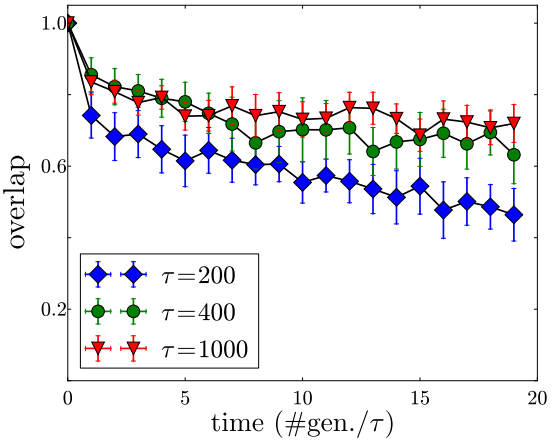


FIG. 4. (Color online) Overlap between functionally significant couplings ($\delta\phi_{ij} > 0.1$) between a system at generation t_0 and a system at generation $t_0 + t$ along a same evolutionary trajectory as a function of the number of generations (time), counted in number of periods τ (the error bars are standard deviations over 100 simulations). The locations of the sectors shown in Fig. 2(b) are found to be stable over multiple periods of environmental changes.

are coevolving, as shown by a statistical analysis of "multiple sequence alignments" obtained from independent evolutionary trajectories with a common origin (Fig. 5).

As indicated by Fig. 2(b), the smaller the period τ of the fluctuations in selective pressure, the smaller the sector. The temporal structure of past selective objectives is thus encoded geometrically in the couplings. More precisely, we may define the sparsity of a system as the fraction S of its couplings J_{ij} with $\delta\phi_{ij} < 0.1$ [the fraction of nonrepresented couplings in Fig. 2(b)]. Sparsity thus represents the fraction of neutral couplings, whose mutation has only a minor effect on the function. It is represented as a function of the period τ in Fig. 2(c) and as a function of τ and the mutation rate μ in

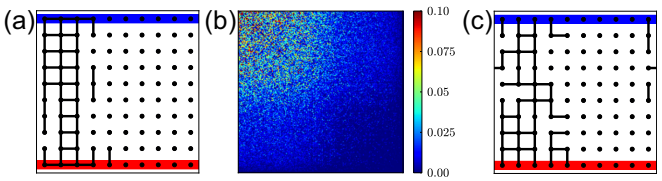


FIG. 5. (Color online) Analysis of coevolution. To generate a data set of phylogenetically related systems analogous to the alignments of protein families used to infer sectors from protein sequences [4], we start from a population obtained by evolution under fluctuating selective pressures and then generate independent trajectories, under the same evolutionary parameters. (a) A system from the initial population, here obtained with $\tau = 200$ and $\mu = 5 \times 10^{-5}$, represented as Fig. 2(b). (b) Matrix of covariance between couplings, $\mathcal{C}_{ij,kl} = |\langle J_{ij} J_{kl}\rangle - \langle J_{ij}\rangle\langle J_{kl}\rangle|$, computed from the results of 100 independent trajectories run over $3\tau = 600$ generations for the same values of the parameters μ, τ; $\langle \ldots \rangle$ denotes an average over the different populations (the absolute value is taken to treat equivalently positive and negative covariations). The couplings are ordered based on the first eigenvector (principal component) of the matrix. (c) The top positions along this principal component define a sector, which overlaps with the sector defined in (a) for the original system.

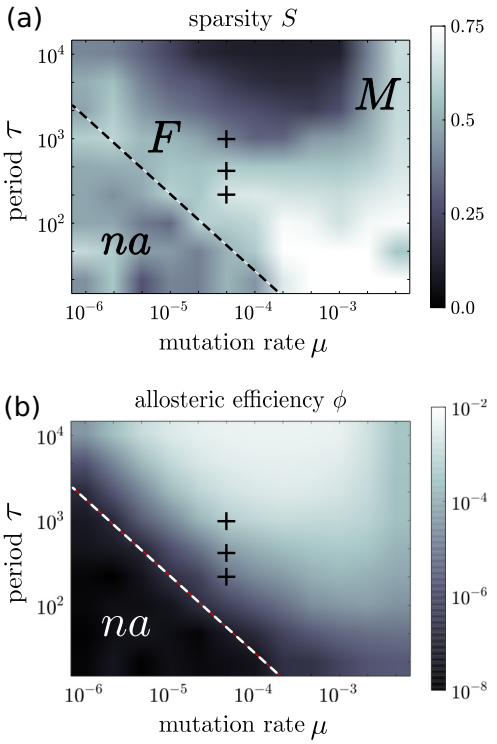


FIG. 6. (Color online) (a) Sparsity of evolved proteins as a function of the mutation rate μ and the period τ of fluctuating selective pressures. Sparsity is defined as the fraction of couplings with $\delta\phi_{ij} < 0.1$ [nonrepresented couplings in Fig. 2(b)]. Below the dashed line, the environmental fluctuations are too fast for the population to follow them, and the systems are nonadapted (region *na*). Sparse systems are found in two ranges of parameters: for intermediate values of $\mu\tau$, where they are driven by fluctuating selective pressures (region F, including the three systems of Fig. 2 indicated by crosses), and for high values of μ, where they are driven by a large mutational load (region M). (b) Fitness of the population as a function of μ and τ; the axis and the red line are the same as in (a). The dashed line ($\phi = 10^{-7}$) corresponds to the typical maximal value of allosteric efficiency in populations of $P = 500$ random systems. Below this line, the populations may be considered as nonadapted.

Fig. 6(a). For low enough mutation rates (see below), it scales with $\mu\tau$, the number of mutations per period; more precisely, it scales with $\mu\tau P$, the total number of mutations in a population of size P (Fig. 7).

Sparsity arises at the expense of instantaneous fitness, here defined by the allosteric efficiency ϕ [Fig. 6(b)], but it favors the "evolvability" [18] of the population, which can be quantified as the fraction of random mutations conferring a noticeable fitness advantage following a change of selective pressure: $\mathcal{E}(J|h') = \langle\theta(\delta\phi^*_{ij} > \epsilon)\rangle_{ij,*}$, where $\langle.\rangle_{ij,*}$ is an average over the pairs ij and over the possible values of J_{ij}, $\theta(x) = 1$ if $x > 0$, 0 otherwise, and ϵ is an arbitrary cutoff [$\epsilon = 0.2$ in Fig. 8(a)]. The evolution of sparsity also implies an increased mutational "robustness" [32], defined as the fraction of mutations that do not affect noticeably the fitness: $\mathcal{R}(J|h) = \langle\theta(\delta\phi^*_{ij} < \epsilon')\rangle_{ij,*}$ where ϵ' is again an arbitrary cutoff [$\epsilon' = 0.01$ in Fig. 8(b)]. The definition of $\mathcal{E}(J|h')$ differ from the definition of $\mathcal{R}(J|h)$ by the field h',

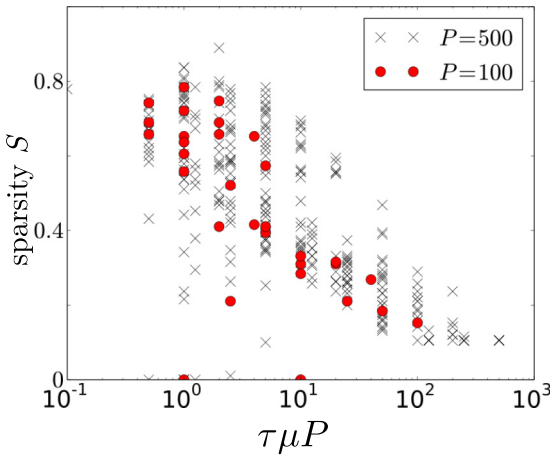


FIG. 7. (Color online) Sparsity as a function of the scaling variable $\tau\mu P$ for systems obtained from evolutionary dynamics with different values of the period τ of environmental changes and mutation rate μ (in the range of values shown in Fig. 6) and for two population sizes, $P = 100, 500$.

which is distinct from the field h in which the system most recently evolved: when considering an environment alternating periodically between two values $h^{(1)}$ and $h^{(2)}$, we thus take the systems at the end of a period of constant selective pressure under $h^{(1)}$ and define robustness as $\mathcal{R}(J|h^{(1)})$ and evolvability as $\mathcal{E}(J|h^{(2)})$.

The period τ is not the only feature of the environmental fluctuations that affects the size of a sector: so does the diversity of these fluctuations. For a given τ, the sparsity thus decreases with the sequence similarity between the two alternating modulators [Fig. 9(a)]. But while the similarity between successive modulators is determining, their exact sequence is not: replacing the sequences $m = (+1, \ldots, +1)$ and $m = (-1, \ldots, -1)$ by arbitrary sequences of ± 1, or even imposing new randomly chosen modulators at each period, does not affect significantly the outcome. This observation

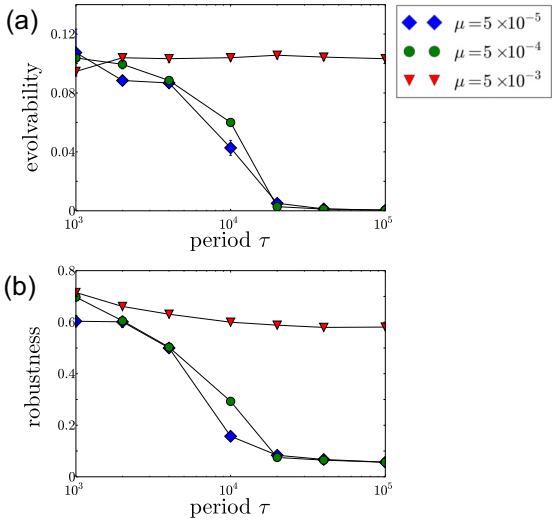


FIG. 8. (Color online) (a) Robustness of evolved systems as a function of the period τ of environmental changes. (b) Evolvability. The error bars are standard deviations over 10 simulations.

illustrates a capacity of "generalization" [33]: the sparse systems, which are more prompt to readapt to a modulator previously encountered in their history, are as prompt to adapt to a modulator never encountered.

Another factor can induce the formation of a sector: a large mutational load. While for small mutations rates μ the sparsity is controlled by the dimensionless parameters $\mu\tau$, for large mutation rates it is controlled by μ nearly independently of τ [Fig. 6(a)]. The critical value of the mutation rate, $\mu_c \sim N^{-1}$, corresponds to the "error threshold" for a system of size N (here the total number of links ij), i.e., to the maximal mutation rate at which a system of this size can faithfully replicate [34]. For $\mu > \mu_c$, the systems thus evolve a sector of size $\sim(\mu N)^{-1}$, which is the largest size allowed by the mutational load.

B. Localization in sequence space

Functional proteins represent only a tenuous subset of all potential proteins [35]. In our model, we find that within this subset, proteins with a sparse sector are themselves rare: typical systems with a given fitness ϕ are significantly less sparse than systems with the same fitness but resulting from an evolution in fluctuating environments [Fig. 10(a)]. This observation implies that sparsity in the evolved systems is not just a consequence of the fitness being curbed by the environmental fluctuations. The sparse systems are, furthermore, not distributed randomly in sequence space, but localized in evolvable regions of this space: they are at shorter mutational distance to solutions to alternative selective pressures, where

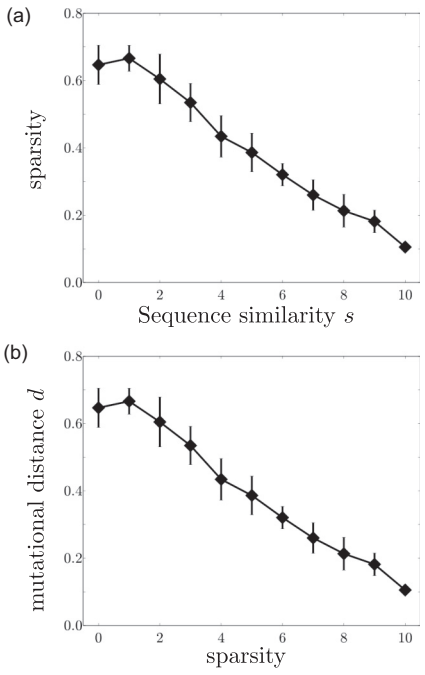


FIG. 9. (a) Sparsity as a function of the sequence similarity s between the two alternative modulator sequences, for $\mu = 10^{-5}$ and $\tau = 100$ (mean and standard deviation over 10 simulations). (b) Minimal number of point mutations necessary to adapt to a new random modulator as a function of sparsity. The simulations are obtained with different values of $\tau \in [10, \ldots, 5000]$ and $\mu \in [5 \times 10^{-3}, \ldots, 2 \times 10^{-6}]$ corresponding to adapted populations ($\phi > 10^{-7}$); the results are averages over five random modulators.

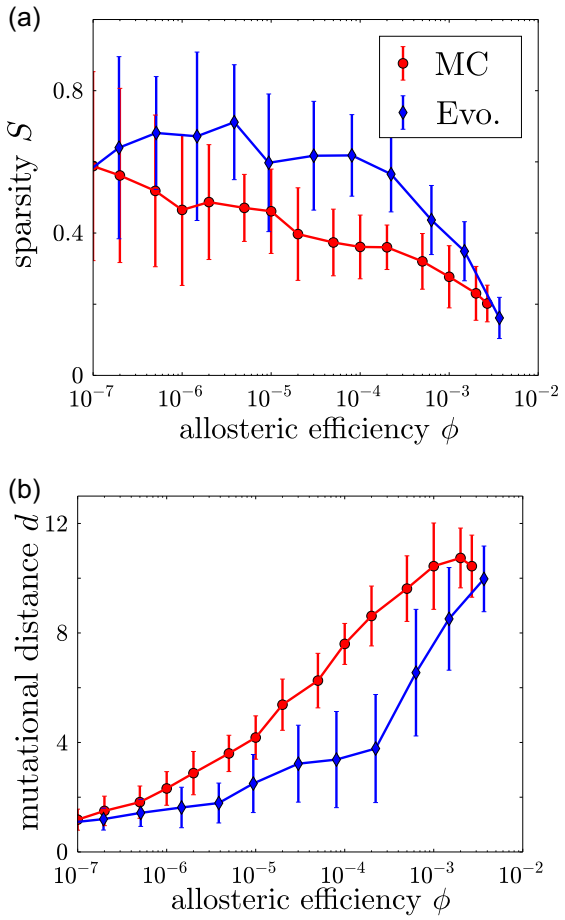


FIG. 10. (Color online) (a) Sparsity as a function of allosteric efficiency (fitness) for evolved systems obtained with various $\mu < 10^{-2}$ and $\tau > 10$ (blue diamond) and for typical systems with same fitness obtained by Monte Carlo sampling (red circle). (b) Distance to a system with different function (i.e., allostery induced by a modulator m' different from the one m for which the system was last selected), measured by the minimal number of beneficial mutations needed to reach an equivalent allosteric efficiency after the change $m \to m'$, for evolved systems (blue diamond) and typical systems (red circle). Systems evolved in a fluctuating environment are thus atypical amongst systems with equivalent fitness value for being sparser and closer to solutions to new selective challenges.

the distance to an alternative selective pressure $h' \neq h$ is estimated as the number of steps necessary for a hill-climbing algorithm, whereby a single coupling J_{ij} can be changed at each step, to reach a fitness value $\phi(J'|h')$ at least equivalent to the initial value $\phi(J|h)$ [Fig. 10(b)]. This phenomenon of localization is generic and has been illustrated previously in the context of fitness landscapes defined on small, schematic sequence spaces [36,37].

Our model, however, displays two features absent from simpler models. First, it relates the topology of the fitness landscape, defined in sequence space, to the geometry of the functional constraints, defined in real space: gradients in fitness thus corresponds to sector positions where adaptive mutations occur, while plateaus in fitness corresponds to positions out of the sector where mutations are almost neutrals. Second, as is typical to high-dimensional spaces, the results are partly nonintuitive: a system localized between two alternating fitness peaks is *ipso facto* localized near a large family of related fitness peaks [Fig. 10(b)], a feature that underlies the faculty of generalization [33], or "promiscuity" [38], previously noted.

C. Modularity

The concentration of functional constraints may take different geometrical forms depending on the structure of the evolutionary fluctuations. In particular, distinct quasi-independent sectors may evolve instead of a single connected sector. A combinatorial process for generating new variations, involving, for instance, gene duplications, recombination events, and/or horizontal transfers, has been shown to produce modular organizations [12]. Such combinatorial variations may explain the modular organization of proteins into domains, which are subsequences of consecutive amino acids, but cannot easily account for the presence of multiple quasi-independent sectors distributed along the sequence of a single domain [4]. A scenario implicating modularly varying selective pressures provides an alternative explanation, as previously illustrated in a range of different models [13,33,39].

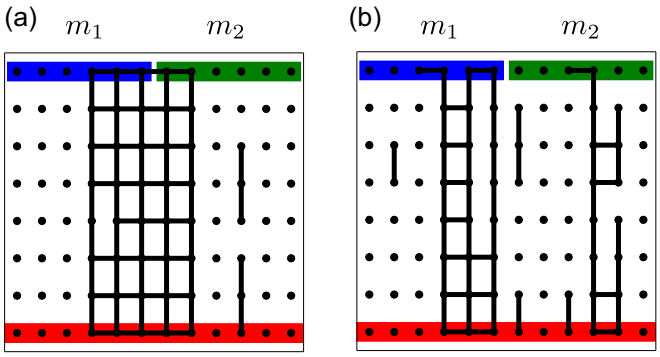
Consistently with these past works, we find that a modularly varying environment favors the emergence of two distinct sectors in an extension of our model where allostery involves two modulators. In this model, the two modulators m_1, m_2 can bind at two distinct regulatory sites (Fig. 11), and selection is for preferential binding of the ligand ℓ in the presence of at least one of them (nonexclusive OR). This corresponds to selecting with a fitness $\phi = \min(\phi_1,\phi_2)$, where the allosteric efficiencies ϕ_1 and ϕ_2 are defined by Eq. (8) with, respectively, $m = (m_1,0)$ and $m = (0,m_2)$.

When both the sequences of m_1 and m_2 fluctuate in time, evolution stochastically generates one of two possible outcomes: systems with a single sector, as in Fig. 11(a), or with two separate sectors, as in Fig. 11(b). The probability to obtain two sectors depends on the structure of the fluctuations (besides the size of the structure): it is significantly larger when m_1 and m_2 change modularly, i.e., one at a time, compared to when they change nonmodularly, i.e., simultaneously (table in Fig. 11).

We note that, in contrast with previous models reporting similar effects [13,33,39], sparsity is not enforced in the definition of the fitness, but obtained as a result of evolution. We also find that a rugged fitness landscape is not necessary for modularity to emerge spontaneously [16]: in our model, solutions are indeed always accessible by hill-climbing with one-step mutations [Fig. 10(b)].

IV. DISCUSSION

Interpreting the information contained in the sequence of a protein requires considering, in addition to the biophysical properties of the protein, its evolutionary history. Our simple model of protein evolution thus demonstrates how a basic feature of proteins, the spatial organization of their residues least tolerant to mutations, may be controlled by past fluctuations of selective pressure or high mutation



Proba. of modularity:

	$\tau = 100$	$\tau = 200$
m_1, m_2 vary together	0.03	0.10
m_1, m_2 vary one at a time	0.82	0.51

FIG. 11. (Color online) Typical outcomes for a variant of the model where the regulatory site is partitioned into two subsites, each associated with an independently varying modulator, and where selection is on allostery in the presence of at least one of the two modulators m_1 and/or m_2. Shown are the couplings with $\delta\phi_{ij} > 0.1$ as in Fig. 2(b). (a) A nonmodular system, with a single sector localized at the interface between the two regulatory sites. (b) A modular system, with two distinct sectors. These two systems are the (stable) outcomes of two distinct evolutionary trajectories with same evolutionary parameters ($\tau = 100$, $\mu = 5 \times 10^{-4}$). The difference stems only from stochastic effects. The table indicates the probability to obtain a modular system for two values of the period τ and for m_1,m_2 varying either simultaneously or consecutively [49].

rates. Our conclusions are based on comparing scenarios that differ only in two evolutionary parameters: the period τ of environmental fluctuations and the mutation rate μ. Since a structural concentration of functional constraints arises only for some values of these parameters, it is clearly not a necessary consequence of the definition of our model.

We expect that comparable results hold for other systems where internal variables varying on a short time scale are subject to short range interactions controlled by evolutionary and environmental variables varying on longer time scales. In less idealized systems, including natural proteins, several additional factors may, however, contribute to a concentration of functional constraints.

Irregular structures thus typically contain preferred allosteric paths that tend to reinforce the concentration of functional constraints: with no unique shortest path between its two interfaces, the cylindrical structure allowed us to illustrate the role of evolutionary factors with minimal contribution from structural heterogeneities. Our approach, however, extends to other geometries.

Similarly, our results are robust to variations in the implementation of the evolutionary dynamics (Fig. 12), but alternative choices may reduce or enhance sparsity [40]. In our model, all coupling values are *a priori* equiprobable, showing that such a mutational bias is not required.

Sparsity may also be favored by factors limiting the efficiency of selection. The typically nonlinear relationship between the biophysical properties of a protein and the reproductive rate (fitness) of organisms may thus make the contribution of all couplings unnecessary. Finite population size effects also generically exclude a complete "optimization" of the couplings.

Our model represents an ideal case where, under constant environment, all the couplings may be equivalently involved in the function (the only *a priori* difference being between vertical and horizontal couplings). In the generic case where a uniform distribution of the couplings is intrinsically nonoptimal, evolutionary fluctuations may, nevertheless, control the degree of concentration of functional constraints if they are sufficiently large.

Many extensions of our model are conceivable. Negative selection against undesired modulators and ligands may, for example, allow us to account for the specificity of the interactions. The assumption of a fixed geometry of interactions may also be relaxed to permit a joint treatment of folding and functional constraints, in line with previous studies based on similar simplified protein models [41–43]. Extending our model to account for structural changes and kinetic effects may thus contribute to rationalize the diversity of mechanisms that evolved to cause allostery [44].

Our model is not intended to account quantitatively for the features of natural proteins. Nevertheless, given the typical size $N \sim 10^2$ and mutations rates $\mu \sim 10^{-9}$ per base per generation of current nonviral proteins, we may exclude a scenario based on high mutation rates for explaining the high tolerance of proteins to mutations. On the other hand, estimates of μP based on silent genomic variations within species give $\mu P \sim 10^{-1}$–10^{-3} for a range of organisms [45], where P represents an effective population size. This indicates that relevant time scales of fluctuating selective pressures are of the order of $\tau \sim (\mu P)^{-1} \sim 10$–$1000$ generations; these estimations are crude but lend weight to the plausibility of a scenario based on environmental fluctuations. Differences of variability in past selective pressures may thus cause different proteins to have fundamentally different architectures of functional constraints.

While our limited knowledge of past evolutionary history prevents us from testing quantitatively these ideas with natural proteins, progress in the field of directed evolution [46–48] may soon offer us a platform to investigate them by performing experiments of evolution under temporally varying selective pressures.

ACKNOWLEDGMENTS

We thank A. Dawid, D. Hekstra, B. Houchmandzadeh, I. Junier, S. Leibler, C. Nizak, K. Reynolds, A. Raman, and R. Ranganathan for discussions and comments. This work was supported by ANR grant CoevolInterProt.

APPENDIX

1. Different modulators

In the main text, we present results when alternating between two opposite modulator sequences, $m^{(1)} = (+, \ldots, +)$ and $m^{(2)} = (-, \ldots, -)$. Any other choice of two opposite modulators with $m_i^{(1)} = -m_i^{(2)}$ for all $i \in$ act gives identical results as a consequence of the "gauge invariance": $\sigma_i \mapsto -\sigma_i \Leftrightarrow J_{ij} \mapsto -J_{ij}\ \forall j$. When alternating between two modulators

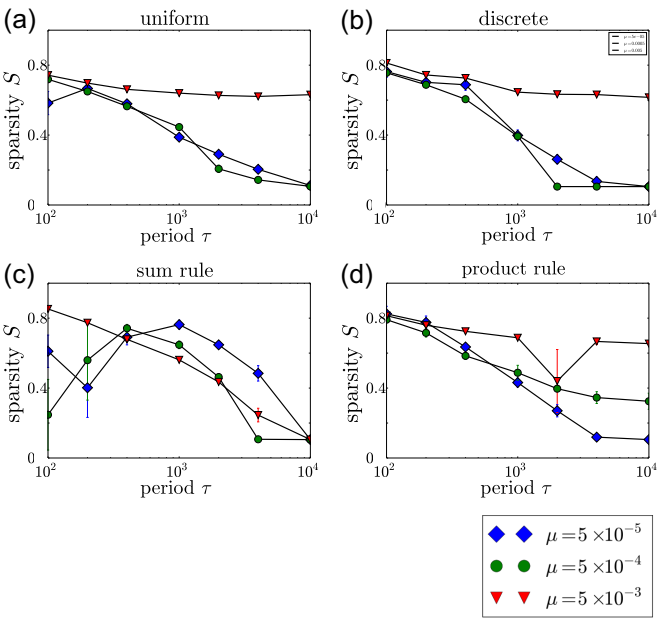


FIG. 12. (Color online) Sparsity as a function of the time scale τ of environmental changes for systems that evolved subject to different mutational processes: (a) Identical to Fig. 2(c): the J_{ij} are mutated to a random value uniformly distributed in $[-1,1]$, independently of their previous value. (b) The J_{ij} are drawn from a finite set of discrete values. (c) Sum rule with variance $\sigma_s^2 = 0.2$. (d) Product rule with variance $\sigma_p^2 = 1.6$. In each case, the sparsity tends to decrease with increasing values of the period τ of the environmental changes.

$m^{(1)}$ and $m^{(2)}$ with $m_i^{(1)} = \pm 1$, $m_i^{(2)} = \pm 1$ and sequence similarity $s = \sum_i \delta(m_i^1, m_i^2)$, the sparsity is commensurate with this measure of similarity [Fig. 9(a)], thus interpolating between the case $s = M = 10$, which is equivalent to a constant environment, and the case $s = 0$, which corresponds to opposite modulators.

Sparsity implies a closer mutational distance to solutions to selective pressures previously encountered in evolutionary history [Fig. 10(b)], but also to new (although related) selective pressures, as illustrated in Fig. 9(b) where evolved systems are challenged with random sequences of the modulator.

2. Alternative mutational processes

The results presented in the main text are obtained with memoryless mutations, consisting in drawing a new value for J_{ij} uniformly at random in $[-1,1]$, independently of its previous value. Among other possible choices, we may consider (i) discrete couplings, taken at random in a finite set of values, $\pm\{0,0.01,0.02,0.05,0.1,0.2,0.5,1\}$; (ii) a sum rule, where each mutation adds to the current value a normally distributed random variable: $J_{ij} \to J_{ij}^* = J_{ij} + \mathcal{N}(0,\sigma_s^2)$; and (iii) a product rule, where each mutation multiplies the current value by a Gaussian variable: $J_{ij} \to J_{ij}^* = J_{ij} \times \mathcal{N}(0,\sigma_p^2)$. We implemented these rules by mapping values $J_{ij}^* > 1$ to $J_{ij}^* = 1$ and values $J_{ij}^* < -1$ to $J_{ij}^* = -1$, to ensure that the couplings remain bounded. The results are presented in Fig. 12, showing that our conclusions are robust with respect to the mutational process.

[1] J. U. Bowie, J. F. Reidhaar-Olson, W. A. Lim, and R. Sauer, Deciphering the message in protein sequences: Tolerance to amino acid substitutions, Science **247**, 1306 (1990).

[2] S. W. Lockless and R. Ranganathan, Evolutionarily conserved pathways of energetic connectivity in protein families, Science **286**, 295 (1999).

[3] R. N. McLaughlin, Jr., F. J. Poelwijk, A. Raman, W. S. Gosal, and R. Ranganathan, The spatial architecture of protein function and adaptation, Nature (London) **491**, 138 (2012).

[4] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, Protein sectors: Evolutionary units of three-dimensional structure, Cell **138**, 774 (2009).

[5] F. C. Peterson, R. R. Penkert, B. F. Volkman, and K. E. Prehoda, Cdc42 regulates the Par-6 PDZ domain through an allosteric CRIB-PDZ transition, Mol. Cell **13**, 665 (2004).

[6] J. Monod, J. P. Changeux, and F. Jacob, Allosteric proteins and cellular control systems, J. Mol. Biol. **6**, 306 (1963).

[7] G. M. Süel, S. W. Lockless, M. A. Wall, and R. Ranganathan, Evolutionarily conserved networks of residues mediate allosteric communication in proteins, Nat. Struct. Biol. **10**, 59 (2002).

[8] R. G. Smock, O. Rivoire, W. P. Russ, J. F. Swain, S. Leibler, R. Ranganathan, and L. M. Gierasch, An interdomain sector mediating allostery in Hsp70 molecular chaperones, Mol. Syst. Biol. **6**, 1 (2010).

[9] K. A. Reynolds, R. N. McLaughlin, and R. Ranganathan, Hot Spots for Allosteric Regulation on Protein Surfaces, Cell **147**, 1564 (2011).

[10] L. Hartwell, J. Hopfield, S. Leibler, and A. Murray, From molecular to modular cell biology, Nature (London) **402**, C47 (1999).

[11] GP. Wagner, M. Pavlicev, and J. Cheverud, The road to modularity, Nat. Rev. Gen. **8**, 921 (2007).

[12] R. V. Solé and P. Fernández, Modularity "for free" in genome architecture?, arXiv:q-bio/0312032.

[13] N. Kashtan and U. Alon, Spontaneous evolution of modularity and network motifs, Proc. Natl. Acad. Sci. USA **102**, 13773 (2005).

[14] C. Branden and J. Tooze, *Introduction to Protein Structure* (Garland Publishing, New York, 1999).

[15] C. Vogel, M. Bashton, N. D. Kerrison, C. Chothia, and S. A. Teichmann, Structure, function and evolution of multidomain proteins, Curr. Opin. Struct. Biol. **14**, 208 (2004).

[16] J. Sun and M. W. Deem, Spontaneous emergence of modularity in a model of evolving individuals, Phys. Rev. Lett. **99**, 228107 (2007).

[17] A. Wagner, *Robustness and Evolvability in Living Systems* (Princeton University Press, Princeton, 2005).

[18] G. P. Wagner and L. Altenberg, Complex adaptations and the evolution of evolvability, Evolution **50**, 967 (1996).

[19] L. W. Ancel and W. Fontana, Plasticity, evolvability, and modularity in RNA, J. Exp. Zool. **288**, 242 (2000).

[20] J. A. Draghi, T. L. Parsons, G. P. Wagner, and J. B. Plotkin, Mutational robustness can facilitate adaptation, Nature (London) **463**, 353 (2010).

[21] R. D. Leclerc, Survival of the sparsest: robust gene networks are parsimonious, Mol. Syst. Biol. **4**, 213 (2008).
[22] H. S. Chan and E. Bornberg-Bauer, Perspectives on protein evolution from simple exact models, Appl. Bioinformatics **1**, 121 (2002).
[23] J. D. Bryngelson and P. G. Wolynes, Spin glasses and the statistical mechanics of protein folding, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).
[24] T. H. Berlin and M. Kac, The spherical model of a ferromagnet, Phys. Rev. **86**, 821 (1952).
[25] I. Bahar, A. R. Atilgan, and B. Erman, Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential, Folding Design **2**, 173 (1997).
[26] D. M. Taverna and R. A. Goldstein, Why are proteins so robust to site mutations?, J. Mol. Biol. **315**, 479 (2002).
[27] G. Tiana, B. E. Shakhnovich, N. V. Dokholyan, and E. I. Shakhnovich, Imprint of evolution on protein structures, Proc. Natl. Acad. Sci. USA **101**, 2846 (2004).
[28] M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[29] I. Bahar, T. R. Lezon, L.-W. Yang, and E. Eyal, Global dynamics of proteins: Bridging between structure and function, Annu. Rev. Biophys. **39**, 23 (2010).
[30] P. Leff, The two-state model of receptor activation, Trends Pharm. Sci. **16**, 89 (1995).
[31] M. Mitchell, *An Introduction to Genetics Algorithms* (MIT Press, Cambridge, MA, 1999).
[32] E. van Nimwegen, J. P. Crutchfield, and M. Huynen, Neutral evolution of mutational robustness, Proc. Natl. Acad. Sci. USA **96**, 9716 (1999).
[33] M. Parter, N. Kashtan, and U. Alon, Facilitated variation: How evolution learns from past environments to generalize to new environments, PLoS Comp. Biol. **4**, e1000206 (2008).
[34] M. Eigen and P. Schuster, *The Hypercycle: A Principle of Natural Self-organization* (Springer, New York, 1979).
[35] A. D. Keefe and J. W. Szostak, Functional proteins from a random-sequence library, Nature (London) **410**, 715 (2001).
[36] L. A. Meyers, F. D. Ancel, and M. Lachmann, Evolution of genetic potential, PLoS Comp. Biol. **1**, e32 (2005).
[37] E. Kussell, S. Leibler, and A. Grosberg, Polymer-population mapping and localization in the space of phenotypes, Phys. Rev. Lett. **97**, 068101 (2006).
[38] O. Khersonsky and D. S. Tawfik, Enzyme promiscuity: A. mechanistic and evolutionary perspective, Ann. Rev. Biochem. **79**, 471 (2010).
[39] N. Kashtan, A. E. Mayo, T. Kalisky, and U. Alon, An Analytically Solvable Model for Rapid Evolution of Modular Structure, PLoS Comp. Biol. **5**, e1000355 (2009).
[40] T. Friedlander, A. E. Mayo, T. Tlusty, and U. Alon, Mutation Rules and the Evolution of Sparseness and Modularity in Biological Systems, PLoS ONE **8**, e70444 (2013).
[41] J. D. Hirst, The evolutionary landscape of functional model proteins, Protein Eng. **12**, 721 (1999).
[42] P. D. Williams, D. D. Pollock, and R. A. Goldstein, Evolution of functionality in lattice proteins, J. Mol. Graph. Model. **19**, 150 (2001).
[43] J. D. Bloom, C. O. Wilke, F. H. Arnold, and C. Adami, Stability and the evolvability of function in a model protein, Biophys. J. **86**, 2758 (2004).
[44] H. N. Motlagh, J. O. Wrabl, J. Li, and V. J. Hilser, The ensemble nature of allostery, Nature (London) **508**, 331 (2014).
[45] M. Lynch, The origins of eukaryotic gene structure, Mol. Biol. Evol. **23**, 450 (2006).
[46] P. A. Romero and F. H. Arnold, Exploring protein fitness landscapes by directed evolution, Nat. Rev. Mol. Cell Biol. **10**, 866 (2009).
[47] K. M. Esvelt, J. C. Carlson, and D. R. Liu, A. system for the continuous directed evolution of biomolecules, Nature (London) **472**, 499 (2011).
[48] A. Fallah-Araghi, J.-C. Baret, M. Ryckelynck, and A. D. Griffiths, A. completely in vitro ultrahigh-throughput droplet-based microfluidic screening system for protein engineering and directed evolution, Lab on a Chip **12**, 882 (2012).
[49] For the statistics shown in the table of Fig. 11, a system is considered as modular if removing the couplings below m_1 (by setting the couplings to 0) leads to a ϕ_2 within 80% of the original ϕ and removing those below m_2 to a ϕ_1 within 80% of the original ϕ; this definition is consistent with a classification based on visual inspection of networks as shown in Fig. 11; it is somewhat arbitrary, but the trends shown in the table of Fig. 11 are not.