Article

Evolution of Specificity in Protein-Protein Interactions

Orit Peleg,[1] Jeong-Mo Choi,[2] and Eugene I. Shakhnovich[2,*]
[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts; and [2]Department of Chemistry and Chemical Biology, Harvard University, Cambridge Massachusetts

ABSTRACT Hub proteins are proteins that maintain promiscuous molecular recognition. Because they are reported to play essential roles in cellular control, there has been a special interest in the study of their structural and functional properties, yet the mechanisms by which they evolve to maintain functional interactions are poorly understood. By combining biophysical simulations of coarse-grained proteins and analysis of proteins-complex crystallographic structures, we seek to elucidate those mechanisms. We focus on two types of hub proteins: Multi hubs, which interact with their partners through different interfaces, and Singlish hubs, which do so through a single interface. We show that loss of structural stability is required for the evolution of protein-protein-interaction (PPI) networks, and it is more profound in Singlish hub systems. In addition, different ratios of hydrophobic to electrostatic interfacial amino acids are shown to support distinct network topologies (i.e., Singlish and Multi systems), and therefore underlie a fundamental design principle of PPI in a crowded environment. We argue that the physical nature of hydrophobic and electrostatic interactions, in particular, their favoring of either same-type interactions (hydrophobic-hydrophobic), or opposite-type interactions (negatively-positively charged) plays a key role in maintaining the network topology while allowing the protein amino acid sequence to evolve.

INTRODUCTION

Proteins that maintain promiscuous molecular recognition, i.e., the ability to maintain functional interactions with multiple partners (1), represent hubs in protein-protein-interaction (PPI) networks. Those highly connected proteins in the cell PPI network are known to play essential roles in cellular control (2–4), and many of them are encoded by essential genes. Therefore, there has been a special interest in the study of the structural and functional properties of hubs that differentiate them from nonhubs (2,5). These differences are manifested in 1), structural properties of the interfaces, and 2), their thermodynamic stability, as further described below.

The identification of structural properties of hubs is based on gene expression patterns and localization of proteins within the cell, as defined by Han et al. (2,3). The authors identified two classes of hubs. The first class contains hubs that interact with all their partners at the same time and in the same space (referred to as Party hubs), and the second consists of hubs that interact with their partners at different times or locations (referred to as Date hubs). The differences between different types of hubs could be manifested in molecular properties (structure and sequence) of Party- and Date-hub proteins. Indeed, based on crystallographic data, Kim et al. (8,9) found two classes of hubs: Multi hubs, which interact with their partners through different interfaces, and Single or Singlish hubs, which interact with most of their partners through one interface. Therefore Singlish hubs cannot interact simultaneously with all their partners, and they generally tend to behave as Date hubs (10).

The next line of evidence highlighting the differences between hubs and nonhubs is based on structural disorder. A hypothetical folding-upon-binding mechanism to provide functional promiscuity in PPIs posits that hub proteins are disordered but get folded upon binding to a partner, potentially acquiring different tertiary structures with different interaction (11,12). Therefore, the study of hub proteins and their interaction partners is relevant for the understanding of the particular role that disorder might play in multifunctional molecular recognition. Bioinformatics analysis provided an initial support to the disorder hypothesis by showing that Singlish hubs and their partners have a higher level of predicted disorder than do Multi hubs, and higher than the proteome average (9). In addition, It has been shown that surfaces of some hubs are enriched in charged and polar amino acids and depleted of hydrophobic content (13–16), a signature of disordered proteins (17). However, despite significant efforts, a clear understanding, at the molecular level, of the physical mechanisms that use hydrophobicity and electrostatics to provide functional promiscuity of hubs in a PPI network remain elusive.

Recently, we developed a multiscale microscopic biophysics-based model to study evolution of functional and nonfunctional PPIs within a simple, yet nontrivial, functional PPI network (18,19). The analysis highlighted an intrinsic conflict in the evolution of hub proteins between

Submitted May 1, 2014, and accepted for publication August 1, 2014.
*Correspondence: shakhnovich@chemistry.harvard.edu
Editor: David Eliezer.

0006-3495/14/10/1686/11 $2.00
http://dx.doi.org/10.1016/j.bpj.2014.08.004

the requirement to maintain multiple functional interactions and the need to avoid ever more abundant nonfunctional ones (20–23). The evolutionary compromise between these factors was achieved in the model (19) by modulating the intracellular abundance of hub proteins and simultaneously decreasing their surface hydrophobic content. Here, we employed the modified biophysics-based model, along with bioinformatics analysis and simple arguments, to study physical principles of evolutionary design of different types of functional PPI hubs. We provide reasoning as to why proteins belonging to Singlish hub systems are more prone to disorder than those belonging to Multi hub systems (9). We found that PPIs of hubs of different types are stabilized by different kinds of physical interactions (e.g., electrostatic, hydrophobic, etc.), both in the biophysics-based model and in the analysis of protein crystallographic databases (Structural Interaction Network (SIN) database (8,9,24)). Finally, we use combinatoric arguments to show that the physical nature of hydrophobic and electrostatic interactions, in particular, their favoring of either same-type (hydrophobic-hydrophobic) or opposite-type interactions (negatively-positively charged), plays a key role in maintaining the network topology while allowing the protein amino acid sequence to evolve. For clarity, we include a flowchart in Fig. 1, depicting our workflow and how results from the lattice-model proteins, SIN database, and combinatoric model are utilized.

METHODS

Lattice protein simulations

We build on a recent biophysics-based multiscale model of evolution wherein simple-lattice-model globular proteins are encoded in genomes of model cells. These genomes are subject to mutations, and the corresponding amino acid sequences determine the precise structure, stability, and interactions of the proteins involved (19,25). In previous work, these simulations have been successful in drawing a direct link between the genotype-phenotype relationship of an organism and several evolved biophysical properties of proteins in its cells, such as their stability, abundance, and PPI strength.

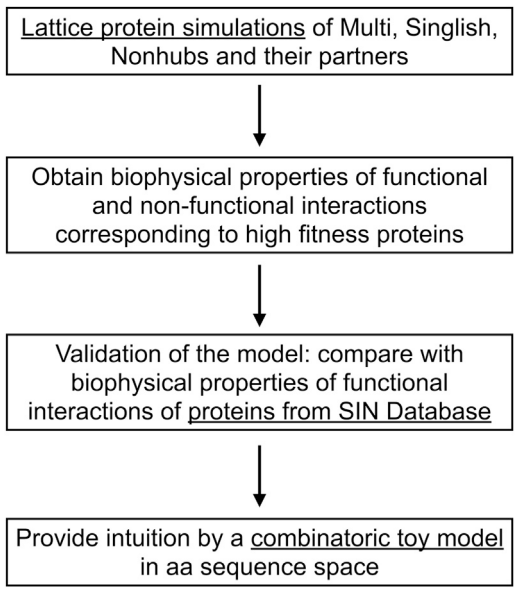

FIGURE 1 Flowchart depicting our workflow, and highlighting how results from each description level (lattice model proteins, SIN database, and combinatoric toy model) are utilized.

Each of our model proteins consists of 27 amino acid residues that fold into $3 \times 3 \times 3$ cubic lattice conformations (Fig. 2 *A*). This 27-mer lattice model protein has 103,346 maximally compact conformations (26). Following Heo et al. (19), for computational efficiency, we use a subset of 10,000 randomly selected conformations as our ensemble. Only amino acids occupying neighboring sites on the lattice can interact, and the interaction energy depends on amino acid types according to the Miyazawa-Jernigan (MJ) potential (27), both for intra- and intermolecular interactions. We calculate the Boltzmann probability of folding to a native state, P^i_{nat}, for each protein:

$$P^i_{\text{nat}} = \frac{e^{-E^i_0/T}}{\sum_{k=1}^{10{,}000} e^{-E^i_k/T}}, \tag{1}$$

where E^i_0 is the energy of the most stable, or native, conformation out of 10,000 conformations and T is the temperature in arbitrary units. The probability of folding, P^i_{nat}, is therefore a proxy for the degree of disorder of protein i. We impose the condition that the native conformation is the minimum energy conformation in the conformational ensemble. Mutations that lead to violation of this condition are considered lethal: every protein in the model proteome is deemed essential, so that its failure to fold deprives the cell of an essential function, causing the lethal phenotype. We model the PPIs with a rigid docking scheme. Six faces of a cubic lattice provide six possible interaction surfaces, and there are four possible directions (which correspond to rotational degrees of freedom) to dock two lattice proteins through two interaction surfaces. Hence, in total, there exist $6 \times 6 \times 4 = 144$ docking modes for a binary protein complex. The Boltzmann probability of interaction in a functional binding mode between proteins i and j is

$$P^{ij}_{\text{int}} = \frac{e^{-E^{ij}_f/T}}{\sum_{k=1}^{144} e^{-E^{ij}_k/T}}, \tag{2}$$

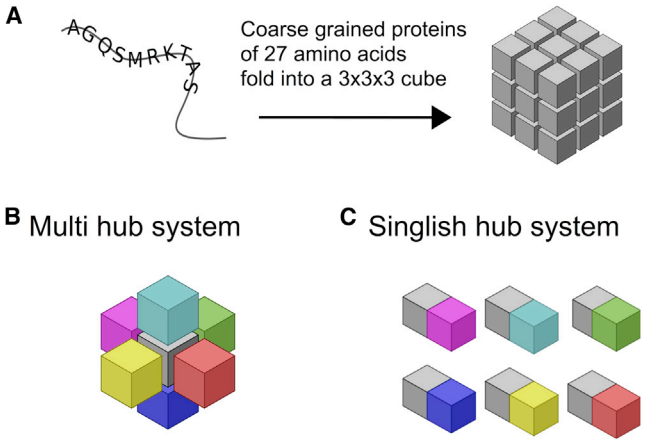

FIGURE 2 (*A*) Schematic representation of the lattice protein model. (*B* and *C*) Model proteins in the system, including the hub protein (*gray*) and up to six interaction partners (*colors*). For the Multi hub system (*B*), each face of the hub is designated to interact with a specific binding partner; For the Singlish hub system (*C*), only a single face is chosen as an interaction interface with all the partners. To see this figure in color, go online.

where E_f^{ij} is the interaction energy of a functional binding mode (defined below), and E_k^{ij} are the interaction energies for all 144 docking modes. We assume that each protein in the cell folds with a two-state folding kinetics,

$$U_i \underset{k_u^i}{\overset{k_f^i}{\rightleftharpoons}} F_i, \tag{3}$$

where U_i denotes unfolded states and F_i the unique native state. k_f^i and k_u^i are the folding and unfolding rate constants, respectively. The steady-state solution is

$$F_i = \frac{k_f^i}{k_u^i} U_i = \frac{P_{\text{nat}}^i}{1 - P_{\text{nat}}^i} U_i. \tag{4}$$

We use the law of mass action (LMA) as described in Heo et al. (25) to calculate all possible functional and nonfunctional interactions between all proteins with folded conformations. We solve the LMA equations numerically (25) to obtain the concentrations of the complexes, with a permitted error of 10^{-15} for each protein concentration. Once the concentrations of protein complexes are obtained, we can define the concentration of functional complexes. The functional concentrations are

$$G_{i,j}^{\text{f}} = \left[F_i F_j\right] \times P_{\text{int}}^{ij}, \tag{5}$$

where $[F_iF_j]$ are the concentrations of dimers formed by folded proteins i and j in any configuration. We consider both Multi and Singlish hubs (Fig. 2, *B* and *C*, respectively). For a Multi hub, each face of the hub is designated to functionally interact with a specific binding partner. For a Singlish hub, a single face is chosen as a functional interaction surface available to all partners, and interactions with different partners are physically mutually exclusive and thus cannot occur simultaneously. We limit the number of partner proteins, N_p, to 6, i.e., the maximal proteome size (including the hub) is 7. Nonfunctional concentrations are defined as

$$G_i^{\text{nf}} = C_i - \sum_{j=1}^{N_p+1} G_{i,j}^{\text{f}}, \tag{6}$$

where C_i is the total concentration of protein i. The fitness of each cell in the population is given by the cell division rate,

$$b = b_0 \frac{\prod_{i=2}^{N_p+1} G_{\text{hub},i}}{1 + \alpha\left(\sum_{i=1}^{N_p+1} C_i - C_0\right)^2}, \tag{7}$$

where b_0 is a constant parameter chosen at the beginning of simulations to scale the rate, and thus the timescale, of evolution, and α scales the overexpression penalty. To limit protein overexpression, we set the parameters $\alpha = 500$ and $C_0 = (N_p + 1) \times 0.1$, as the initial protein concentration was chosen to be $C = 0.1$. We use the Gillespie algorithm (28) with fixed population size ($N = 100$). Upon cell division, a mother cell gives birth to a daughter cell. To keep the population size constant, a newborn cell replaces a randomly chosen cell in the population. Upon replication, both the mother and daughter cells are subjected to either a mutational event with constant rate ($\mu = 10^{-3}$/gene/replication) or to a change in the expression level of one protein in the cell. This protein is chosen randomly, and its expression level changes with a constant rate ($r = 0.01$/cell division) such that its concentration in the new cell is obtained from the old one as $C^{\text{new}} = C^{\text{old}} + \epsilon$, where ϵ is a Gaussian random number with zero mean and variance 0.1. All simulations ran for 6×10^6 generations and were considered to be in a steady state when C_i, P_{nat}^i, and P_{int}^{ij} reached a plateau value. The random seed of each simulation is initialized with a unique value. The results described below are averaged over 50 different random simulations.

To calculate the evolutionary rates of proteins in the simulation, we count the number of fourfold, twofold, and nondegenerate sites (corresponding to four, two, and one amino acid representation in nucleotide space), and use the formula of Hartl and Clark ((29), page 340):

$$\begin{aligned} N_{\text{s}}^t &= (\text{fourfold degenerate sites}) \\ &\quad + 1/3 \times (\text{twofold degenerate sites}) \\ N_{\text{a}}^t &= (\text{nondegenerate sites}) \\ &\quad + 2/3 \times (\text{twofold degenerate sites}), \end{aligned} \tag{8}$$

where N_{s}^t is the number of synonymous sites at generation t, and N_{a}^t is the number of nonsynonymous sites at generation t. Next, we calculate

$$\begin{aligned} dN(t) &= 2A_{\text{s}}^t / \left(N_{\text{s}}^0 + N_{\text{s}}^t\right) \\ dS(t) &= 2A_{\text{a}}^t / \left(N_{\text{a}}^0 + N_{\text{a}}^t\right), \end{aligned} \tag{9}$$

where N_{s}^0 and N_{a}^0 are the numbers of synonymous and nonsynonymous sites, respectively, of the initial sequence. A_{s}^t and A_{a}^t are the numbers of synonymous and nonsynonymous amino acids, respectively, at generation t. The evolutionary rate is then defined to be dN/dS.

SIN database analysis

To compare the lattice proteins to natural proteins, we used the SIN database (8,9,24), which identified Multi and Singlish hubs and their available atomistic structures in complex with their partners at the Protein Data Bank (PDB). We used the information available for human proteins, in total 37 Multi, 36 Singlish, and 30 Nonhub interactions (see Table S1 in the Supporting Material for the list of PDB files considered). Nonhub interactions were defined as interactions between proteins that do not belong to a Singlish or Multi hub group. We identified and analyzed the interfaces, and studied the statistics of contacts (two residues are defined to be in contact when any pair of atoms from the two different amino acids are separated by <4 Å). To facilitate comparison with simulations we used the MJ potential to evaluate the contact energies.

Combinatoric toy model

The model consists of three proteins (a hub and two partners), where each protein has two interaction surfaces. The simulation space is one-dimensional (1D) and proteins can interact via a single interface. In total, there are six surfaces in the system, and surfaces can be hydrophobic, hydrophilic, or positively or negatively charged. This corresponds to the reduction of the amino acid pull from 20 to 4. We use an averaged MJ potential for the four types of amino acids, where hydrophobic amino acids were identified as M, F, I, L, V, W, P, and C, negatively charged as D and E, and positively charged as R and K (30) (as done for the lattice-model simulation and SIN database analysis). The model system has a countable sequence space. In particular, there are six surfaces, each of which can be one of four types; thus, there are a total of $4^6 = 4096$ possible configurations.

For each configuration, we solve the LMA equations

$$C_i + C_j \overset{K_{i,j}}{\rightleftharpoons} C_i \times C_j, \tag{10}$$

where C_i and C_j are the concentrations of protein i and protein j, respectively. The initial concentrations were chosen to be $C_{\text{hub}} = 0.5$ and $C_{\text{par}} = 0.1$ (the latter set for both partners). The association constants are defined as

$$K_{i,j} = \frac{[C_i \times C_j]}{[C_i][C_j]} = \sum_{k,m=1}^{3 \text{ or } 4} e^{-E_{k,m}}, \tag{11}$$

where $E_{k,m}$ is the averaged MJ interaction energy mentioned above and k and m are the particular bound surfaces of proteins i and j, respectively. For interaction between two different proteins, there are four modes of interaction. For interaction between two identical proteins, there are three modes of interaction due to the 1D symmetry in the system. The functional concentrations are defined as

$$\begin{aligned} G_{\text{hub,par1}} &= [C_{\text{hub}} \times C_{\text{par1}}] \frac{e^{-E_{\text{pair1}}}}{\Sigma_{k,m=1}^{4} e^{-E_{\text{k,m}}}} \\ G_{\text{hub,par2}} &= [C_{\text{hub}} \times C_{\text{par2}}] \frac{e^{-E_{\text{pair2}}}}{\Sigma_{k,m=1}^{4} e^{-E_{\text{k,m}}}}. \end{aligned} \tag{12}$$

The functional pair varies between Multi, Singlish, and Nonhub systems. For Multi hubs, each surface of the hub is designated for a different partner. For Singlish hubs, the same surface of the hub interacts with both partners. For Nonhubs, we define a single functional interaction, whereas the second partner is required to remain in a monomeric form. For hub systems, the fitness of the configuration is

$$f = G_{\text{hub,par1}} \times G_{\text{hub,par2}}, \tag{13}$$

whereas for Nonhub systems, the fitness of the configuration is

$$f = G_{\text{hub,par1}} \tag{14}$$

RESULTS

Lattice protein simulations

Even though the simulated-lattice-protein Multi, Singlish, and Nonhub systems considered here differ in their functional interaction networks and numbers of partners, some trends are similar in most systems. 1), Shortly after the beginning of the simulations, the concentration of the hub proteins increases such that $C^{\text{hub}} = \sum_i C_i^{\text{partner}}$ (after ~ 10^4 generations; see Fig. 3 *A*). 2), Hub and partner proteins eventually evolve a stable structure (P_{nat} increases), whereas some proteins experience a sharp drop in P_{nat} beforehand (see Fig. 3 *B*). 3), The nonfunctional concentration, G_i^{nf}, of the hub increases at first due to the increase in hub concentration, but when the hub concentration, C, stabilizes, the G_i^{nf} of the hub proteins decreases. The G_i^{nf} of partners decreases monotonically throughout the simulation (see Fig. 3 *C*). 4), The hub-partner pairs develop functional surfaces with increasing probability of functional interaction, P_{int} (see Fig. 3 *D*).

Effect of number of partners

One particular goal was to understand the evolutionary processes that dominate changes in the thermodynamic stability of hub proteins and their level of disorder. In both Multi and Singlish hub systems, hub and partner proteins evolve a stable structure (P_{nat} increases). Although the P_{nat} of hubs tends to increase monotonically, those of partner proteins experience a drop before maximal stability is reached. We studied the effect of number of partners on the evolution of protein stability, as reflected in P_{nat}. Fig. 4 shows the P_{nat} of hubs

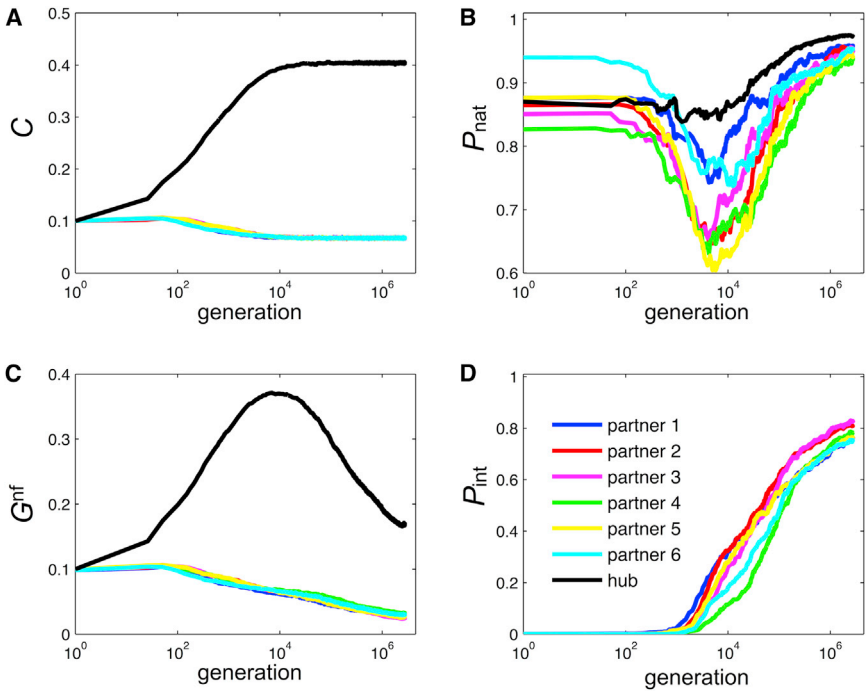

FIGURE 3 Evolution of molecular properties in Singlish hub systems with six interaction partners with regard to concentration, C, of hub (*black*) and partner proteins (*colors*; see legend in panel D) (*A*), thermodynamic stability, P_{nat}^i (*B*), nonfunctional concentration, G_i^{nf} (*C*), and functional interaction probabilities, P_{int}^{ij} (*D*). Data are plotted versus generation, averaged over 50 independent realizations (of different random-number seed), and shown in logarithmic scale for the x axis. The Multi hub system shows similar behavior for these quantities. To see this figure in color, go online.

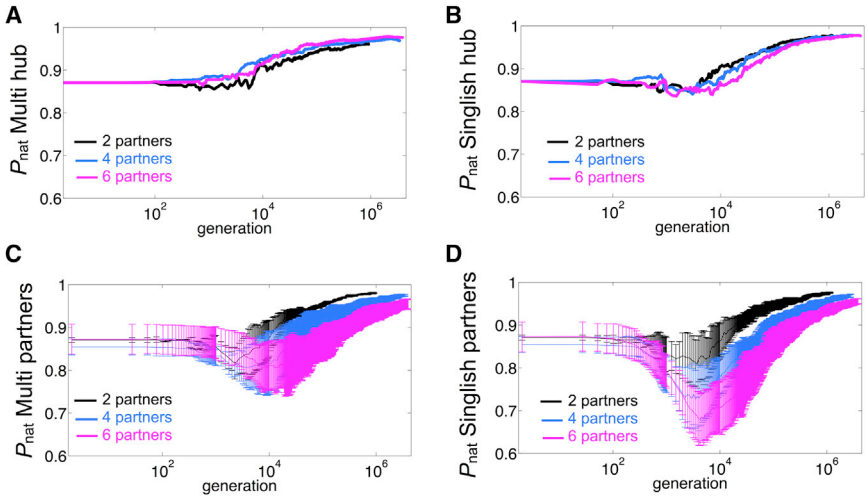

FIGURE 4 Effect of the number of partners on stability, P_{nat}, of proteins as a function of generation for Multi hubs (*A*), Singlish hubs (*B*), partners of Multi hubs (*C*), and partners of Singlish hubs (*D*). Systems with two, four, and six partners are shown in black, blue and magenta, respectively. For partner proteins, P_{nat} is averaged over all partners in the system, and error bars correspond to 1 Standard Deviation (SD). All plotted data (including averaged P_{nat} and SD) are averaged over 50 independent realizations and shown in logarithmic scale for the *x* axis. To see this figure in color, go online.

and partners for Multi (Fig. 4, *A* and *C*, respectively) and Singlish hub systems (Fig. 4, *B* and *D*, respectively) and compares systems with two, four, and six partners. The P_{nat} of partners of both Multi and Singlish hub systems drops with respect to its initial value. As the number of partners increases, both the amplitude of the drop and the time necessary for stabilization after reaching the minimal P_{nat} increase. The drop in P_{nat} is more profound for partners of Singlish hubs in comparison to those of Multi hubs. As partners lose stability, they open the sequence space needed to improve the interaction strength, P_{int} (Fig. 5). However, functional interaction surfaces in Multi hub systems evolve more slowly and poorly, as can be seen in Fig. 5 for systems with a maximal partner number of six.

Evolutionary rates

In previous studies, the evolutionary rate of proteins has been shown to correlate negatively with the number of functional interfaces (8), yet a clear understanding of the behavior of the evolutionary rates of hub proteins and partners throughout evolution remains elusive. Therefore, we calculated the evolutionary rates, dN/dS, by counting the number of synonymous and nonsynonymous sites in the nucleotide sequence (see further details and Eq. 9 in Methods).

In Fig. 6, we plot the evolutionary rate for hub proteins (Fig. 6 *A*) and partner proteins (Fig. 6 *B*) as a function of generation. Our results are in agreement with those of Kim et al. (8) and show that the more functional interfaces a protein has, the lower is its evolutionary rate in steady state (once the evolutionary rate reaches a plateau); Singlish hubs, which have a single interaction surface, reach a higher evolutionary rate than do Multi hubs, which have six functional interfaces (Fig. 6 *A*). Partners of both Multi and Singlish hubs reach a similar plateau value (Fig. 6 *B*). Both Singlish hubs and their partners experience an increase in evolutionary rate, ~ 10^4 generations before Multi hubs and their partners.

Biophysics of specificity

Next we turn to the question of how hubs use the hydrophobic and charge interactions to design their surfaces to maximize functional interactions. To that end, we examined the functional energy contributions from hydrophobic and electrostatic contacts once simulations reached a steady state (after 6×10^6 generations). Contacts between amino acid pairs were classified as hydrophobic contacts if both amino acids were hydrophobic (hydrophobic amino acids were identified as M, F, I, L, V, W, P, and C (30)), or as electrostatic contacts if the pair contained a positively and a

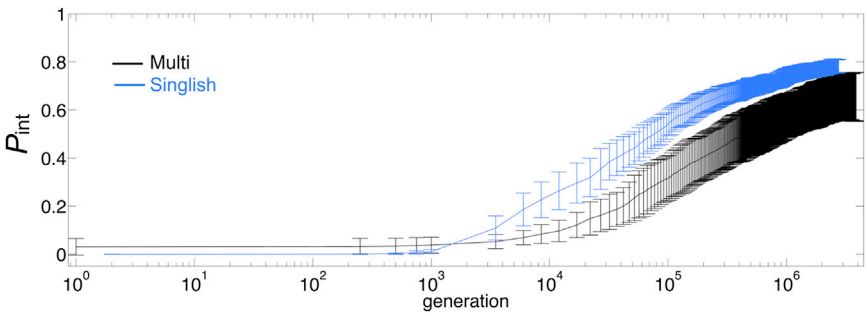

FIGURE 5 Comparing P_{int} for Multi and Singlish hub systems with the maximal number of partners (six). P_{int} is averaged over all partners in the system, and error bars correspond to 1 SD. P_{int} and SD values are averaged over 50 independent realizations and shown in logarithmic scale for the *x* axis. To see this figure in color, go online.

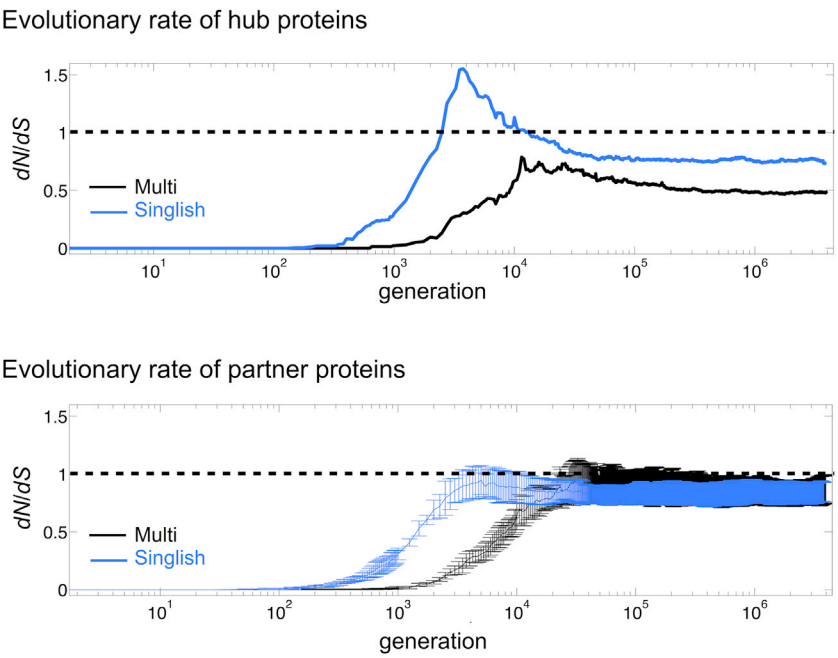

FIGURE 6 Evolutionary rates, dN/dS, versus generation, for hub proteins (A) and partner proteins (B). In B, the data are averaged over all six partners, and error bars correspond to 1 SD. Plotted data are averaged over 50 independent realizations and shown in logarithmic scale for the x axis. To see this figure in color, go online.

negatively charged amino acid (negatively charged amino acids were identified as D and E, and positively charged amino acids as R and K (30)). Values of the hydrophobic and electrostatic energies were normalized by the total interaction energy of their corresponding functional interface. We extracted those energies for 50 different realizations, and we plot the hydrophobic and electrostatic functional energy histograms in Fig. 7 *A* and *B*, respectively. The latter show that Multi hubs use electrostatic contacts more than Singlishs do. In particular, the mean electrostatic contribution is $18.6 \pm 0.1\%$ and $13.2 \pm 0.1\%$ for Multi and Singlish hubs, respectively ($\pm$ values represent the variance calculated over multiple realizations; Kolmogorov-Smirnov (KS) test, $p = 9.6 \times 10^{-18}$). Nonhubs have a lower mean electrostatic contribution $12.7 \pm 0.3\%$ compared to hub systems. The same analysis for hydrophobic contacts shows an opposite trend: mean contributions of $52.6 \pm 0.4\%$ and $66.2 \pm 0.3\%$ for Multi and Singlish hubs, respectively (KS test, $p = 3.9 \times 10^{-32}$). Nonhubs have a higher mean hydrophobic contribution $(69.4 \pm 0.4\%)$ in comparison to hubs.

As the simulations gave us clear insights and predictions regarding the steady-state biophysical properties of hub-protein systems, the next step was to test these predictions for natural proteins. To that end, we used the SIN database (8,9,24), which identified Multi and Singlish hubs and their available atomistic structures in complex with their partners at the PDB (see further details in Methods). As in the simulations analysis, we compared the percentages of hydrophobic-hydrophobic and opposite-charge contacts for both Multi and Singlish hubs (see Fig. 7, *C* and *D*). This analysis verified our prediction that electrostatic contacts are more predominant in Multi hubs and hydrophobic contacts are more predominant in Singlish hubs. The electrostatic contributions calculated using the SIN database are $15.5 \pm 1.1\%$ and $7.8 \pm 0.5\%$ for Multi and Singlish hubs, respectively ($\pm$ values represent variance calculated over multiple PDB files; KS test, $p = 0.001$). Nonhubs have an electrostatic contribution $(8.0 \pm 0.3\%)$ comparable to that of Singlish hubs. The same analysis for hydrophobic contacts shows average contributions of $27.8 \pm 2.1\%$ and $28.4 \pm 2.5\%$ for Multi and Singlish hubs, respectively (KS test indicates a nonsignificant difference between the two). However, Nonhubs have a higher hydrophobic contribution $(38.6 \pm 2.4\%)$ in comparison to hubs (KS test, $p = 0.012$ and $p = 0.004$ for Multi and Singlish hubs, respectively).

Next, we analyzed the nonfunctional interaction energies. Due to the geometrical complexity of the crystallographic 3D structure of proteins, nonfunctional PPI interfaces are not clearly defined, and therefore, analysis of the nonfunctional energies is not feasible for proteins in the SIN database. However, nonfunctional interaction energies are accessible in the lattice-protein simulations, and their energetic contribution is calculated as for the functional ones. Again, we consider both hydrophobic and electrostatic contributions to the interaction energies. As we are interested in the negative design of nonfunctional interactions (i.e., minimizing nonfunctional interactions), we extract information about the following least favorable interactions: hydrophobic with nonhydrophobic (marked as the hydrophobic contribution), and electrostatic interactions between charges with the same sign (marked as the electrostatic contribution).

We plot the nonfunctional energy histograms in Fig. 8, showing the mean energetic contributions of both

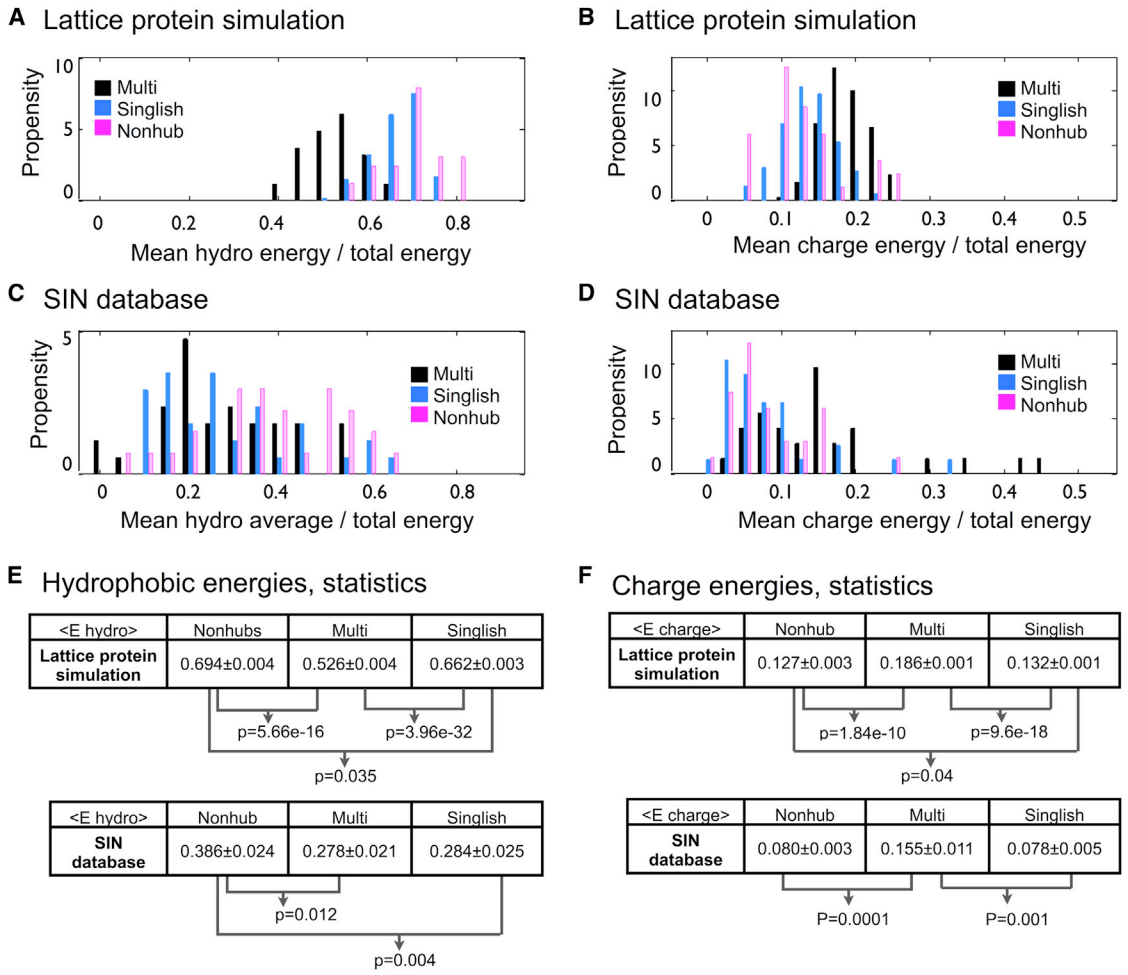

FIGURE 7 Mean interaction energy histograms showing fractions of hydrophobic and electrostatic contributions for lattice protein simulations (*A* and *B*, respectively) and SIN database (*C* and *D*, respectively). For the lattice protein, energies are calculated at the end of the simulations at generation 6×10^6. Multi and Singlish hub systems are shown in black and blue, respectively. Nonhub systems are shown in magenta. Data were binned to generate hydrophobic energy histograms (bin size $dx = 0.05$), and charge energy histograms (bin size $dx = 0.025$). (*E*) Statistics of hydrophobicity histograms expressed as mean ± variance. (*F*) Statistics of electrostatic histograms expressed as mean ± variance. The *p* values were extracted using the KS test. Only significant values ($p<0.05$) are noted. To see this figure in color, go online.

hydrophobic (Fig. 8 *A*) and electrostatic contacts (Fig. 8 *B*) divided by the total energy of the interaction. Similar to the case for functional interactions, Multi hubs use electrostatic contacts in their nonfunctional PPI network more than Singlish hubs and Nonhubs. In particular, the electrostatic contribution is $8.8 \pm 0.004\%$ for Multi hubs, and $7.8 \pm 0.005\%$ and $7.8 \pm 0.003\%$ for Singlish and Nonhubs, respectively (KS test, $p = 7.65 \times 10^{-20}$ and $p = 2.27 \times 10^{-9}$ for Multi-Singlish and Multi-Nonhub, respectively). The electrostatic contribution is not significantly different for Singlish hubs and Nonhubs. Therefore, in comparison to Singlish hubs and Nonhubs, Multi hubs use more electrostatic interactions to both maximize probabilities for functional interactions and minimize probabilities for nonfunctional interactions.

The same analysis for hydrophobic contacts shows that the distributions of Multi hubs, Singlish hubs, and Nonhubs are significantly different (see *p* values in Fig. 8 *C*). Nonhubs use the smallest number of hydrophobic contacts for the nonfunctional interactions, $18.7 \pm 0.005\%$ on average. The average hydrophobic contribution is $25.5 \pm 0.022\%$ for Singlish hubs and $26.9 \pm 0.021\%$ for Multi hubs. The trend showing the average hydrophobic contribution is highest for Multi hubs, intermediate for Singlish hubs, and lowest for Nonhubs is exactly the opposite of their relationship when accounting for functional interactions. Therefore, Multi hubs are highly efficient in negative design of their nonfunctional surfaces for minimizing unfavorable interactions, both hydrophobic and electrostatic.

Combinatoric toy model

Finally, we turn to a simpler model to provide insight into why Multi hubs rely on electrostatics to design their

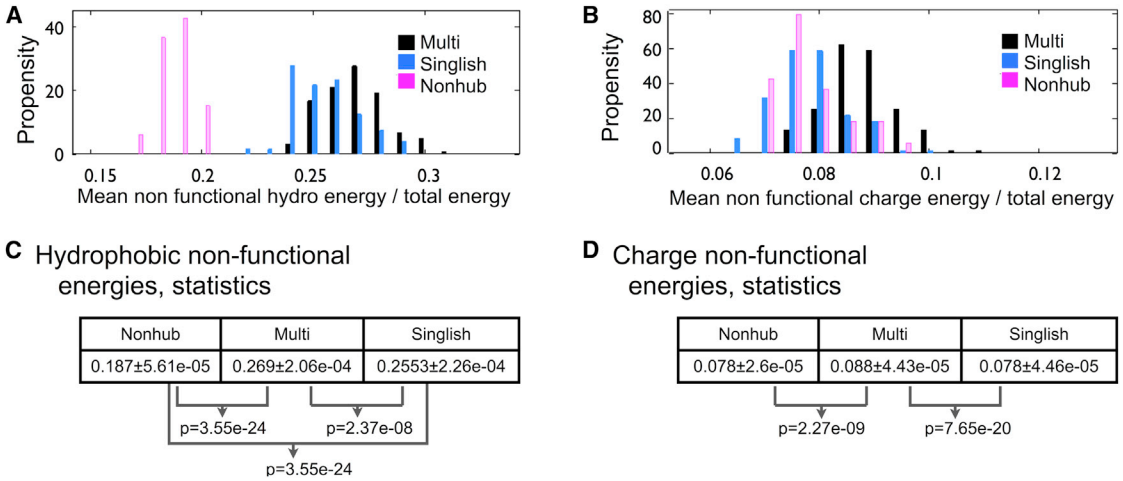

FIGURE 8 Negative design in the multiscale evolutionary model. Nonfunctional interaction energy histograms showing the contribution fraction of hydrophobic interactions (*A*) and electrostatic interactions (*B*) for lattice protein simulations. Multi and Singlish hub systems are shown in black and blue, respectively. Nonhub systems are shown in magenta. Bin sizes used to generate hydrophobic energy histograms and charge energy histograms were $dx = 0.01$ and $dx = 0.005$, respectively. (*C*) Statistics of hydrophobicity histograms expressed as mean ± variance. (*D*) Statistics of electrostatic histograms expressed as mean ± variance. The p values were extracted using the KS test. Only significant values ($p<0.05$) are noted. To see this figure in color, go online.

functional interaction surfaces, whereas Singlish hubs make predominant use of hydrophobic interactions. One possible explanation involves combinatorics, or counting of states, where a state is a certain amino acid sequence along the surfaces of the proteins. In particular, we wish to identify high-fitness sequences for Multi hubs, Singlish hubs, and Nonhubs and to classify them according to their hydrophobicity and charge content. To that end, we created a simplified 1D model of a hub and two partners, where each protein has two surfaces. Each surface is assigned to be one of four types: hydrophobic, positively or negatively charged, or neutral. The fitness (defined in Eqs. 13 and 14) is calculated according to the probability of functional interactions (see Methods for further details). The advantages of the simplified model are 1), the feasibility of sequence space enumeration, and 2), the isolation of surface energetics from thermodynamic stability. The combination of these advantages allowed us to address the fitness-maximization problem as a tiling puzzle, where electrostatic and hydrophobic surfaces are distributed among the proteins.

We start by considering the set of 80 top fitness configurations shown in Fig. 9, *A–C*, for Multi, Singlish, and Nonhub systems. The vertical axis indicates fitness. For each system, functional interactions are indicated in magenta for the top configuration. Hydrophobic surfaces are shown in green, hydrophilic in black, positively charged in red, and negatively charged in blue. We find that the maximal fitness configurations for both Multi and Singlish hub systems contain hydrophobic functional surfaces. For Multi hub systems, the first ~40 configurations have two hydrophobic functional contacts, but the rest of the top configurations have one hydrophobic and one electrostatic functional contact. This is different from the Singlish- and Nonhub systems, where all top configurations have two hydrophobic functional contacts.

Further understanding of the origin of the favoring of electrostatic functional interactions in Multi hub systems can be extracted from the fitness distributions plotted in Fig. 9, *D–F*, where we plot histograms of the number of configurations (in log scale) corresponding to a certain relative fitness (fitness divided by maximal fitness). The configuration count is divided for all hydrophobic functional contacts (*green*), all electrostatic functional contacts (*red*), and one hydrophobic and one electrostatic contact (*blue*, not available for the Nonhub systems, since they only have one functional contact). When fitness decreases, more configurations containing electrostatic functional surfaces become available, most profoundly for Multi hub systems. In situations where the maximal-fitness configuration is not easily approached (for example, due to limitations arising from thermodynamic stability considerations, i.e., stand-alone lattice proteins are known to be most stable when ~30% of their amino acid sequence is hydrophobic (19)), the next available configurations are more likely to contain electrostatic functional interfaces in Multi hub systems in comparison to Singlish-hub systems. This may explain why Multi hubs use more electrostatic than hydrophobic functional interactions and why the reverse is true for Singlish hubs in the lattice protein simulations and in the SIN database (see Biophysics of specificity). This effect can be quantified by the average fitness of the configurations considered, and indeed, the average fitness of configurations involving electrostatic functional interactions for Multi hub systems is $\langle f \rangle = 0.109 \pm 0.001$, which is

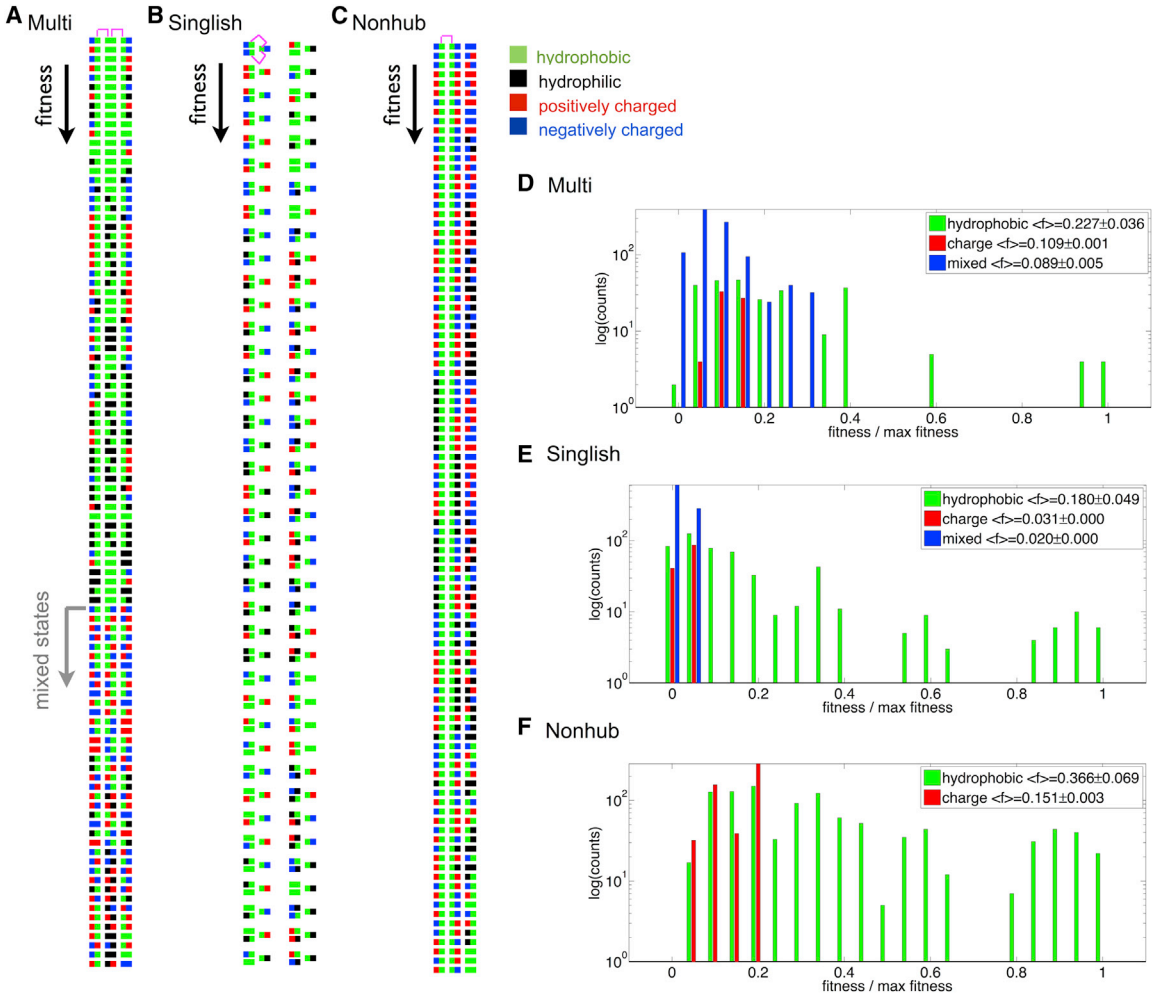

FIGURE 9 Results using a simplified 1D model. (*A–C*) The top 80 configurations for Multi (*A*), Singlish (*B*), and Nonhub systems (*C*). Configurations are arranged vertically according to their fitness (highest fitness at the top). For each system, functional interactions are indicated in magenta for the top configuration. Hydrophobic surfaces are shown in green, hydrophilic in black, positively charged in red, and negatively charged in blue. (*D–F*) Histograms of the number of configurations (in log scale) corresponding to a certain relative fitness (fitness divided by maximal fitness) are shown for Multi (*D*), Singlish (*E*), and Nonhub systems (*F*). The configuration count is divided for all hydrophobic functional contacts (*green*), all electrostatic functional contacts (*red*), and one hydrophobic and one electrostatic contact (*blue*; not available for the Nonhub system). Mean ± variance values of the distributions are indicated in the legend. To see this figure in color, go online.

approximately three times larger than that for Singlish-hub systems, $\langle f \rangle = 0.031 \pm 0.0001$.

DISCUSSION

The findings presented here have implications for the conceptual understanding of the emergence of disorder in PPI networks, aswell as the design of functional interactions within them. We start by providing reasoning as to why proteins belonging to Singlish-hub systems are more prone to disorder than those belonging to Multi hub systems (9). Our simulations show that throughout evolution, partners of Singlish hubs experience a larger drop in stability, P_{nat}, (i.e., they explore more disordered configurations), and a longer time to recover their maximal P_{nat}, in comparison to Multi hub partners (Fig. 4, *C* and *D*). Hubs in the simulations experience a profoundly smaller drop in P_{nat} (Fig. 4, *A* and *B*), despite the bioinformatics analysis predicting a high level of disorder for Singlish hubs as well (9). This discrepancy could be attributed to the fact that in our simulations, proteins are assigned hub or partner exclusively, whereas in nature, hub proteins may become partners in other hub systems, particularly in Singlish-hub systems (9). Singlish-hub partners may utilize the drop in P_{nat} to open up a sequence space necessary for the evolution of functional interactions. Indeed, evolutionary rates of proteins in our simulations (Fig. 6) show a trend similar to that described by Kim et al. (8), who observed that evolutionary rates negatively correlate with the number of functional interfaces. Moreover, our simulations show that proteins in Singlish-hub systems spend more time in functional interaction complexes compared to proteins in Multi hub systems

and therefore evolve a more robust and efficient PPI (Fig. 5).

In addition, our simulations and analysis of the SIN database reveal two different mechanisms of surface design: whereas Multi hubs rely on electrostatics to design their functional interaction surfaces, Singlish hubs make predominant use of hydrophobic interactions (Fig. 7, *C* and *D*). Despite the simplified nature of the protein structure in the simulations, estimation of the contact energies for both simulated coarse-grained and atomistic real protein structures follow the trend mentioned above (Fig. 7, *A* and *B*) and suggest that our coarse-grained simulations capture the fundamental physics of hub protein systems. It is remarkable that different ratios of hydrophobic to electrostatic interactions of interfacial amino acids maintain distinct network topologies (i.e., Multi and Singlish) and underlie a fundamental design principle of PPI. Furthermore, the PPI takes place in a crowded environment, where functional interactions must overcome nonfunctional ones. Indeed, we show that hubs are highly efficient in designing their nonfunctional surfaces as well, for minimizing energetically favorable yet nonfunctional interactions; hubs use more hydrophobic mismatches (noncompatible amino acid pairs at the hub-partner interface) than do Nonhubs (Fig. 8 *A*), and Multi hubs use more electrostatic mismatches than do Singlish hubs and Nonhubs do (Fig. 8 *B*).

To further understand the origin of the two design mechanisms, we turned to a simplified 1D system of a hub and two partners, where each protein has two surfaces. Each surface is modeled as one of four types: hydrophobic, positively or negatively charged, or neutral. The model provided evidence that the different design mechanisms of Multi and Singlish hub systems arise from a limited number of configurations (i.e., sequences of surface types) in the system. As for the lattice protein simulations, high fitness requires both maximizing probabilities of functional PPIs and minimizing probabilities of nonfunctional PPIs. After enumerating all possible configurations and ranking them according to their fitness, we revealed that there are more high-fitness configurations employing charge in Multi hub systems than in Singlish hub systems (Fig. 9). This combinatoric reasoning implies that the physical nature of hydrophobic and electrostatic interactions, in particular their favoring of either same-type interactions (hydrophobic-hydrophobic) or opposite-type interactions (negatively-positively charged), plays a key role in maintaining the network topology while allowing the protein amino acid sequence to evolve.

CONCLUSIONS

In conclusion, our results highlight the origin of thermodynamic instabilities and several biophysical steady-state properties of hub protein systems. Although our findings reveal the fundamental design principles of these systems, there are opportunities to extend the scope of this work, focusing, for example, on hierarchical systems composed of several hub networks and testing our predictions by experiments in vitro and possibly in vivo.

SUPPORTING MATERIAL

One table with details regarding the PPIs considered in the analysis of the SIN database is available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(14)00841-8.

This work was financially supported by Swiss National Science Foundation (SNSF) grant no. PBEZP3_140130 4 (O.P.) and National Science Foundation grant MCB-1243837 (E.S.).

REFERENCES

1. Mosca, R., R. A. Pache, and P. Aloy. 2012. The role of structural disorder in the rewiring of protein interactions through evolution. *Mol. Cell. Proteomics.* 11: M111.014969.
2. Han, J. D. J., N. Bertin, …, M. Vidal. 2004. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature.* 430:88–93.
3. Krogan, N. J., G. Cagney, …, J. F. Greenblatt. 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature.* 440:637–643.
4. Jeong, H., S. P. Mason, …, Z. N. Oltvai. 2001. Lethality and centrality in protein networks. *Nature.* 411:41–42.
5. Zotenko, E., J. Mestre, …, T. M. Przytycka. 2008. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLOS Comput. Biol.* 4:e1000140.
6. Reference deleted in proof.
7. Reference deleted in proof.
8. Kim, P. M., L. J. Lu, …, M. B. Gerstein. 2006. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science.* 314:1938–1941.
9. Kim, P. M., A. Sboner, …, M. Gerstein. 2008. The role of disorder in interaction networks: a structural analysis. *Mol. Syst. Biol.* 4:179. http://dx.doi.org/10.1038/msb.2008.1016.
10. Andorf, C. M., V. Honavar, and T. Z. Sen. 2013. Predicting the binding patterns of hub proteins: a study using yeast protein interaction networks. *PLoS ONE.* 8:e56833.
11. Dunker, A. K., M. S. Cortese, …, V. N. Uversky. 2005. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.* 272:5129–5148.
12. Dyson, H. J., and P. E. Wright. 2002. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* 12:54–60.
13. Patil, A., and H. Nakamura. 2005. Filtering high-throughput protein-protein interaction data using a combination of genomic features. *BMC Bioinformatics.* 6:100.
14. Patil, A., and H. Nakamura. 2006. Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. *FEBS Lett.* 580:2041–2045.
15. Higurashi, M., T. Ishida, and K. Kinoshita. 2008. Identification of transient hub proteins and the possible structural basis for their multiple interactions. *Protein Sci.* 17:72–78.
16. Higurashi, M., T. Ishida, and K. Kinoshita. 2009. PiSite: a database of protein interaction sites using multiple binding states in the PDB. *Nucleic Acids Res.* 37:D360–D364. http://dx.doi.org/10.1093/Nar/Gkn659.
17. Uversky, V. N., J. R. Gillespie, and A. L. Fink. 2000. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins.* 41:415–427.

18. Cetinbaş, M., and E. I. Shakhnovich. 2013. Catalysis of protein folding by chaperones accelerates evolutionary dynamics in adapting cell populations. *PLOS Comput. Biol.* 9:e1003269.

19. Heo, M., S. Maslov, and E. Shakhnovich. 2011. Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. *Proc. Natl. Acad. Sci. USA.* 108:4258–4263.

20. Zhang, J., S. Maslov, and E. I. Shakhnovich. 2008. Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size. *Mol. Syst. Biol.* 4:210. http://dx.doi.org/10.1038/msb.2008.1048.

21. Deeds, E. J., O. Ashenberg, …, E. I. Shakhnovich. 2007. Robust protein protein interactions in crowded cellular environments. *Proc. Natl. Acad. Sci. USA.* 104:14952–14957.

22. Johnson, M. E., and G. Hummer. 2011. Nonspecific binding limits the number of proteins in a cell and shapes their interaction networks. *Proc. Natl. Acad. Sci. USA.* 108:603–608.

23. Levy, E. D., S. De, and S. A. Teichmann. 2012. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proc. Natl. Acad. Sci. USA.* 109:20461–20466.

24. Bhardwaj, N., A. Abyzov, …, M. B. Gerstein. 2011. Integration of protein motions with molecular networks reveals different mechanisms for permanent and transient interactions. *Protein Sci.* 20:1745–1754.

25. Heo, M., L. Kang, and E. I. Shakhnovich. 2009. Emergence of species in evolutionary "simulated annealing". *Proc. Natl. Acad. Sci. USA.* 106:1869–1874.

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

27. Miyazawa, S., and R. L. Jernigan. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 256:623–644.

28. Gillespie, D. T. 2007. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.* 58:35–55.

29. Hartl, D. L., and A. G. Clark. 2007. Principles of Population Genetics. Sinauer Associates, Sunderland, MA.

30. Berezovsky, I. N., K. B. Zeldovich, and E. I. Shakhnovich. 2007. Positive and negative design in stability and thermal adaptation of natural proteins. *PLOS Comput. Biol.* 3:e52.