Supporting Information

Sikosek et al. 10.1073/pnas.1115620109

SI Text

The Hydrophobic-Polar (HP) Lattice Protein Model. The sequence-to-structure mapping used in our investigation was based on the two-dimensional (2D) HP model of protein folding (1). The HP model is a highly coarse-grained approach that aims to account for hydrophobic interaction, which is a main driving force in protein folding (2). Protein sequences in the model consist of only two types of residues—hydrophobic (H) or polar (P)—and protein chains are configured as self-avoiding walks (SAWs) on a 2D square lattice. The only type of favorable interactions in the model is between two sequentially nonadjacent hydrophobic residues that are spatially nearest neighbors on the lattice. Each such hydrophobic-hydrophobic (HH) contact is assigned an energy $\varepsilon(< 0)$. The total energy of a conformation with h HH contacts is equal to εh (1, 3–5).

In the present context, a gene is equivalent to a model protein sequence of H and P residues. Different SAWs (those that are not related by rigid rotations or inversions) correspond to different protein chain conformations. The terms "conformation" and "structure" are used interchangeably in our discussion, whereas a "state" refers to a set of one or more conformations. For any given HP sequence, the density of states $g(h)$ is the number of conformations as a function of the number of HH contacts, h. The number of conformations with the maximum number, h_{N}, of HH contacts achievable by a given sequence is denoted by g and is referred to as the ground-state degeneracy of the sequence (4); i.e., $g \equiv g(h_{\mathrm{N}})$. In view of the experimental observation that many natural globular proteins adopt an essentially unique native structure under folding conditions, we take $g = 1$ HP sequences as models for globular proteins (4).

As in our previous evolutionary studies (6–10), data for chain length $n = 18$ were analyzed in this work. The usage of short 2D HP sequences to model the general behavioral trends of much longer three-dimensional (3D) protein chains is justified in at least two respects. First, because of the importance of hydrophobicity in protein folding, an important geometric factor in protein energetics is the ratio between the number of surface-exposed and buried residues. In this regard, the exterior/interior ratios of folded 2D conformations with chain length $n \sim 16$ are comparable to the surface/core-volume ratio of 3D globular proteins with approximately 150 residues (5), which correspond roughly to the chain lengths of many enzymes that have been studied by biophysical methods. Second, the nonrandom distribution of hydrophobic residues along natural protein sequences (11) is well rationalized by the corresponding nonrandom distribution in short, 2D $g = 1$ HP sequences of length $n = 18$ (12). Reminiscent of the architecture of some natural proteins, certain short $g = 1$ sequences in the 2D HP model are comprised of autonomous folding units (7, 13). Taken together, these considerations support our working assumption that the 2D HP model is a useful theoretical construct for capturing salient features of the sequence-to-structure mapping of real proteins (6, 7, 14), despite the model's insufficiency for a full account of the thermodynamics and kinetics of cooperative protein folding (15, 16). Beside the 2D HP model, other lattice protein models have also been employed to study evolution (17–20). Earlier advances in evolutionary applications of simple lattice protein models were reviewed in refs. 13 and 21; a detailed assessment of the merits and limitations of HP and other lattice models in the study of protein folding can also be found in ref. 14.

For the $n = 18$ 2D HP model used here, the mapping from all $2^{18} = 262{,}144$ possible HP sequences onto all 5,808,335 possible conformations, without any restrictions such as maximal compactness (13), was obtained by exact (exhaustive) enumeration (6). This mapping provides the density of states $g(h)$ for each of the sequences. Let X be the set (ensemble) of all possible protein conformations (structures) in the model. In accordance with Boltzmann statistics, for any given gene (sequence) i with density of states $g(h)$, the fractional population $\Phi(X_l, i)$ of any conformation $X_l \in X$ is equal to

$$\Phi(X_l, i) = e^{-\epsilon h_l / k_{\mathrm{B}} T} \Big/ \sum_h g(h) e^{-\epsilon h / k_{\mathrm{B}} T}, \qquad \textbf{[S1]}$$

where h_l is the number of hydrophobic contacts in X_l, k_{B} is the Boltzmann constant, T is absolute temperature, and the summation is over all possible h values in X (6). The fractional population $\Phi(X_l, i)$ is a measure of stability $(0 < \Phi(X_l, i) < 1)$ and is expected to be positively correlated with the concentration C_l of functional protein molecules that are folded as X_l (see below). We used $-\varepsilon / k_{\mathrm{B}} T = 5.0$ for the folding condition in our computation to ensure that an overwhelming majority of the sequences in the neutral nets we studied have dominant ground-state (native) populations. When $-\varepsilon / k_{\mathrm{B}} T = 5.0$, the median fractional populations of the native structures in the neutral networks A and B in Fig. 1*A* of the main text are, respectively, 0.87 and 0.73. Among the 48 $g = 1$ sequences in network A, the minimum fractional native population is 0.68. Among the 20 $g = 1$ sequences in neutral network B, only two sequences have their fractional native populations fall below 0.5.

A neutral network in our model is a collection of protein sequences that encode for the same native structure and are interconnected by single-point mutations. Because genes in our model are the protein sequences themselves and mutations are performed directly on the protein sequences, all mutations in our model are nonsynonymous. Multifunctionality is modeled by a given set of beneficial structures X^{b}; and the evolutionary role of multifunctionality is assessed by considering the neutral networks of each of the structures in X^{b}. The present investigation focuses on X^{b} with two structures.

The largest neutral network in the $n = 18$ 2D HP model consists of 48 $g = 1$ sequences (nodes) encoding uniquely for the same native structure X_A (conformation drawn in blue at the top of Fig. 1*A* of the main text). This network, referred to as "A" here, was featured in several previous publications (6, 7, 9, 14). Neutral network A is directly connected to another neutral network "B" with native structure X_B (drawn in red at the top of Fig. 1*A* of the main text) and consists of 20 $g = 1$ sequences. The direct connectivity between the neutral networks means that a single mutation can convert a $g = 1$ sequence in network A to a $g = 1$ sequence in network B. In addition to the $g = 1$ sequences, multiply-degenerate $(g > 1)$ sequences with $g = 2, 3, 4, 5,$ and 6 are also included in the network pairs in Fig. 1*A* as long as their ground-state conformations contain either X_A or X_B, or both. There are 84 and 40 such multiply-degenerate sequences in the extended neutral networks (6) for X_A and X_B, respectively. Among them are seven bridge sequences that contain both X_A and X_B in their native states. These bridge sequences provide equal stability to X_A and X_B and belong to both extended neutral networks.

Using $S_{X^{\mathrm{b}}}$ to denote the set of all single genes belonging to either or both of two interconnected neutral networks for $X^{\mathrm{b}} = \{X_A, X_B\}$ and ω to denote the total number of such genes, $D_{X^{\mathrm{b}}} \equiv S_{X^{\mathrm{b}}} \times S_{X^{\mathrm{b}}}$ is the set of all ω^2 gene pairs (i, j) where $i, j \in S_{X^{\mathrm{b}}}$. Gene pairs are ordered, i.e., $(i, j) \neq (j, i)$, to reflect ordering of the genes along the DNA. A gene pair can arise from a single gene by gene duplication or from an existing gene pair by point mutation. In our model, $G_{X^{\mathrm{b}}} \equiv S_{X^{\mathrm{b}}} \cup D_{X^{\mathrm{b}}}$ is the set of all $\Omega = \omega + \omega^2$ genotypes relevant to X^{b}. In total, there are $\omega = 48 + 20 + 84 + 40 - 7 = 185$ single genes and $\Omega = 34{,}410$ genotypes in the network pair A and B in Fig. 1*A* of the main text. In our evolution dynamics simulations (see below), all sequences in the $g \leq 6$ extended networks of a given network pair (e.g., A and B) are considered to be viable (with nonzero fitness values), whereas any mutation that takes a sequence outside the given network pair and/or results in a $g > 6$ sequence is considered lethal (zero fitness for the mutant).

Because both X_A and X_B are targets for selection in our fitness function for the network pair A and B (see below), bridge sequences in the overlapping region of the two extended neutral networks for A and B have high fitness in our model. Other (nonbridge) $g > 1$ genes have relatively low fractional populations of one or both (beneficial) target structures, and thus low fitness. The upper bound for the fractional population of any ground-state conformation is $1/g$ because all ground-state conformations are equally probable in the present model. The most stable bridge sequence β_{AB} (most stable with respect to X_A and X_B) has only X_A and X_B as ground-state conformations, whereas other bridge sequences have additional nonbeneficial ground-state conformations. Under our simulation condition with $-\varepsilon/k_{\mathrm{B}}T = 5.0$, the fractional populations of X_A and X_B for the most stable bridge sequence are equal to $\Phi(X_A, \beta_{AB}) = \Phi(X_B, \beta_{AB}) = 0.444$. In general, a higher stability for such a bridge gene (i.e., higher Φ for the target structures) means increased fitness and thus enhanced relevance of the bridge gene for EAC. Under the same $-\varepsilon/k_{\mathrm{B}}T = 5.0$ condition, the fractional populations of X_A and X_B for the prototype π_A of network A are $\Phi(X_A, \pi_A) = 0.998$ and $\Phi(X_B, \pi_A) = 4.532 \times 10^{-5}$, respectively. The corresponding fractional populations for the prototype π_B of network B are $\Phi(X_A, \pi_B) = 6.438 \times 10^{-3}$ and $\Phi(X_B, \pi_B) = 0.955$. The prototype sequences π_A and π_B are shown in Fig. S1*A*. Additional information about the network pair A and B and other network pairs used in the present study are provided in Table S1. A discussion of bridge ("switch") sequences in the 2D HP model was first given in ref. 22. Another example of a neutral network that is connected by a bridge sequence to neutral network A can be found in figure 16.7 of ref. 14.

The Fitness Function. The fitness function $W(C_l)$ of a structure X_l in our model depends on the functional concentration C_l (also referred to simply as "concentration" below). For a genotype consisting of a single gene i, we set $C_l(i) = \Phi(X_l, i)$. For a genotype comprising of a pair of genes (i, j), we considered two alternate definitions of $C_l(i, j)$ to compare behaviors in the absence and presence of dosage effect, and use the shorthands $d = 0$ and $d = 1$ to label the two cases, respectively. The rationale and the biological ramifications for considering these two scenarios are outlined in the main text and will be discussed further in this *SI Text* below. The $d = 0$ case assumes that there is no change in functional concentration immediately after gene duplication, at which time $i = j$ and $C_l(i, j) = \Phi(X_l, i) = \Phi(X_l, j)$. When $\Phi(X_l, i) \neq \Phi(X_l, j)$ because of subsequent mutations, we set $C_l(i, j)$ to the larger (max) of the two Φ values. For $d = 1$, concentration is equal to the sum of the fractional populations of the two genes in the same genotype, viz.,

$$C_l(i,j) = \begin{cases} \max[\Phi(X_l, i), \Phi(X_l, j)] & \text{if } d = 0 \\ \Phi(X_l, i) + \Phi(X_l, j) & \text{if } d = 1. \end{cases} \qquad \textbf{[S2]}$$

In the present formulation, the functional concentration for a gene pair in the $d = 1$ scenario is determined by the total Boltzmann population of the beneficial structure contributed by both genes. In contrast, the functional concentration for a gene pair in the $d = 0$ scenario is determined only by the Boltzmann population contributed by the gene that imparts a higher thermodynamic stability on that structure. Therefore, in general, instead of the $d = 1$ functional concentration that treats all X_l population equally, the $d = 0$ functional concentration prescribes higher weights to more stable populations of X_l. A situation in which a $d = 0$-like assignment of functional concentration may apply is when the kinetic stability of a protein (23), i.e., the duration the protein stays in the beneficial structure before transiently adopting another conformation, is important for performing its biological function. Further discussion of this rationale and its ramifications are provided below. Our fitness function $W(C_l)$ is controlled by two parameters θ and τ:

$$W(C_l) = \begin{cases} \theta[1 - (1 - C_l/\theta)^{1/\tau} + (C_l/\theta)^{\tau}]/2 & \text{if } C_l < \theta \\ \theta & \text{if } C_l \geq \theta \end{cases} \qquad \textbf{[S3]}$$

with $\theta \in [0, 1]$ serving as an upper bound on $W(C_l)$ (Fig. S1*B*). As discussed in the main text, θ may be viewed as the selection pressure on C_l (cf. ref. 20) because a small θ means that a low concentration of X_l does not sacrifice fitness. In contrast, when $\theta = 1$, $C_l(i)$ has to approach the highest possible concentration of unity to achieve maximum fitness; any decrease in stability is detrimental. A similar fitness function based on the thermal adaptation of adenylate kinase was derived by Peña et al. (24). τ was not discussed in the main text. It parametrizes the deviation from a linear relationship between fitness and concentration for $C_l \in [0, \theta]$. Depending on τ, this relationship can be convex, linear, or concave (Fig. S1*B*). For example, when $\tau < 1$, changes in promiscuous functions at low concentrations (gray shade in Fig. S1*B*) have a stronger positive impact on fitness whereas changes in native functions at high concentrations have relatively smaller fitness effects. In this manner, τ parameterizes the intrinsic relationship between beneficial function and concentration of a protein structure in a given biological setting. Throughout the main text, $\tau = 1$ was implied.

The total fitness of a genotype k, consisting of either one gene ($k \equiv i$) or two genes ($k \equiv i, j$), is defined as the sum of fitness contributions from each conformation X_l belonging to the set of beneficial structures X^{b}; i.e., $W_k = \sum_{X_l \in X^{\mathrm{b}}} W(C_l)$, where $C_l = C_l(i)$ for $k \equiv i$ and $C_l = C_l(i, j)$ for $k \equiv (i, j)$ (Eq. **S2**). As stated above, our analysis is focused on X^{b} with two target structures ($X^{\mathrm{b}} = \{X_A, X_B\}$, for example).

For the neutral networks A and B in Fig. 1*A* of the main text with 48 and 20 $g = 1$ sequences, respectively, the total number of SUBF pairs is $48 \times 20 \times 2 = 1{,}920$. There are only two sequences in network B with $\Phi < 0.5$ (see above). Thus, when $\theta = 0.5$, the fitness values among SUBF are high (average $= 0.99$, standard deviation $= 0.03$). It also follows that the number of subfunctionalized (SUBF) pairs with suboptimal fitness is $48 \times 2 \times 2 = 192$, and hence the total number of SUBF pairs with optimal fitness used in the analysis in Fig. 3 of the main text is 1,920–192 = 1,728.

Two Fitness Parameters Determine the Degree of Functional Trade-Off in Our Model. Both the selection pressure θ and the intrinsic relationship between function and concentration parametrized by τ in our model can impact on the functional trade-off between two structures. Fig. S1*C* illustrates the general effects of τ and θ on the degree of trade-off and the resulting evolutionary process (neofunctionalization, NEOF or subfunctionalization, SUBF). As a comparison with the results for $\tau = 1, \theta = 0.5$ or 1 presented in the main text and Figs. S2 and S3, results for other (τ, θ) values in Fig. S1*C* are provided in Fig. S4. Evolutionary dynamics simulations confirm the general trend that NEOF follows from a strong trade-off whereas SUBF follows from a weak trade-off (see below).

Evolutionary Dynamics of Genotype Populations Using a Master-Equation Treatment. The present formalism is an adaptation of the master-equation approach in refs. 6 and 7 to incorporate effects of gene duplication. For any genotype k in the set of all genotypes relevant to the set of beneficial structures X^{b} (i.e., $k \in G_{X^{\mathrm{b}}}$), let $P_k(q)$ be the time-dependent probability or fractional population normalized by the total population in $G_{X^{\mathrm{b}}}$, where time q is the number of generations (time steps). The populations at time step q determine those at the next time step $q+1$. For genotypes with a single gene, $k \equiv i \in S_{X^{\mathrm{b}}}$,

$$P_i(q+1) = \left[-(n\mu + \mu_d)P_i(q) + \mu \sum_{r=1}^{A_i} P_{\nu_i(r)}(q) + P_i(q)\right] \frac{\mathcal{N}(q) W_i}{\bar{W}(q)}, \qquad \text{[S4]}$$

where μ is the point mutation rate for any given position along a gene sequence of length n; $\nu_i(r)$, where $r = 1, 2, \ldots$, labels the A_i genes adjacent to i, i.e., those that differ by one point mutation from gene i; and μ_d is the gene duplication rate of converting any genotype with a single gene into a genotype with a pair of identical genes. Network topology is defined by gene adjacencies, which were determined by exact enumeration in our model (13). Starting with population of i at q (last term inside the square brackets in Eq. **S4**), the first term on the left accounts for population loss caused by outgoing mutations and by conversion of the single-gene genotype to a double-gene genotype in one time step, whereas the second term accounts for population gain resulting from incoming mutations during the same time step. $W_i/\bar{W}(q)$ is a reproduction factor that depends on relative fitness, where W_i is the fitness of gene i and $\bar{W}(q) \equiv \sum_{k=1}^{\Omega} P_k(q) W_k$ is the population average of the fitness values of all genotypes in $G_{X^{\mathrm{b}}}$ at time q. $\mathcal{N}(q) = 1/\sum_{k\in G_{X^{\mathrm{b}}}} P_k(q+1)$ normalizes the total population to unity (7) to facilitate comparison of population distributions at different time steps. $\mu = 10^{-3}$ and $\mu_d = 10^{-4}$ were used to obtain the results in Fig. 2 in the main text for sequences of length $n = 18$. For genotypes with two genes, $k \equiv (i, j) \in D_{X^{\mathrm{b}}}$,

$$\begin{aligned} P_{ij}(q+1) = \Bigg[& -2\mu n P_{ij}(q) + \mu \sum_{r=1}^{A_i} P_{\nu_i(r)j}(q) + \mu \sum_{s=1}^{A_j} P_{i\nu_j(s)}(q) \\ & + \mu^2 \sum_{r=1}^{A_i} \sum_{s=1}^{A_j} P_{\nu_i(r)\nu_j(s)}(q) + \delta_{ij}\mu_d P_i(q) \\ & + P_{ij}(q) \Bigg] \frac{\mathcal{N}(q) W_{ij}}{\bar{W}(q)}, \qquad \text{[S5]} \end{aligned}$$

where the first term on the right accounts for population loss due to point mutations in both genes i and j, which have a combined sequence length $2n$. The next three terms account for population gain from point mutations of genes adjacent to i and j that reside either in single-gene genotypes (second and third terms) or in double-gene genotypes (fourth term). When $i = j$, population of (i, i) can also increase because of duplication of i. This gain is accounted for by the fifth term where the Kronecker symbol $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. W_{ij} is the fitness of genotype (i, j) and $W_{ij}/\bar{W}(q)$ is the corresponding reproduction factor. Eqs. **S4** and **S5** thus describe a Markov process of population dynamics for the evolving genotypes that is governed by fitness as well as by network topology. Results in the main text and in Figs. S3–S5*A* and S6 were obtained using the above master-equation approach, a schematic summary of the formulation is provided in Fig. S7.

NEOF and SUBF Follow from Strong and Weak Trade-Off, Respectively. In addition to the results in the main text obtained using model parameter sets $(\tau = 1, \theta = 0.5)$, and $(\tau = 1, \theta = 1)$, we have further explored the general relationship between functional trade-off on the one hand and NEOF versus SUBF on the other by determining the evolutionary dynamics for the other four (τ, θ) parameter sets in Fig. S1*C* using the master-equation approach (Fig. S4). When the selection pressure θ is low, functional trade-off is extremely weak. Consequently, the fitness increase associated with bridge genes can be so strong that gene duplications provide no further advantage, let alone NEOF or SUBF (Fig. S4*A*, for $\tau = 1$, and $\theta = 0.25$). In the case of a weak trade-off but a high selection pressure, SUBF can be an adaptive process even with dosage effect ($d = 1$) because promiscuous functions at low concentrations are highly rewarded (Fig. S4*B*, for $\tau = 0.3$ and $\theta = 1$). When the selection pressure is moderately high and the trade-off becomes less weak, a "hybrid" between NEOF and SUBF ensued, in that bridge pairs are populated but bridge genes are not (Fig. S4*C*, for $\theta = 0.75$ and $\tau = 1$). As expected, when both the selection pressure and trade-off are strong, the result is NEOF (Fig. S4*D*, for $\theta = 1$ and $\tau = 2$). This NEOF outcome is very similar to the $\theta = \tau = 1$ case in the main text.

Network Layout. The network layout ("sequence space") in Fig. 1*A* of the main text was generated by the Fruchterman–Reingold algorithm (25) to facilitate visualization of network topology. In the present construction, an edge is assigned to connect two nodes if and only if the sequences represented by the two nodes differ by a single-point mutation (i.e., they are separated by Hamming distance 1). The Fruchterman–Reingold algorithm was applied to keep edge lengths as similar as possible so that distances along edges reflect roughly the Hamming distances between sequences. In this algorithm, the edges act like springs that can be stretched or compressed by other nodes and edges until a certain equilibrium state is achieved. The algorithm is nondeterministic because initial node placements are random. As a result, final placements of nodes can vary slightly between replicates, but the overall layout properties will be very similar. It should be emphasized that the Fruchterman–Reingold technique was adopted here solely for presentational purposes. It has no bearing on the computational analysis of our model.

Stochastic Monte Carlo Simulations for Finite Populations. The master-equation approach we introduced (Eqs. **S4** and **S5**) describes a deterministic process that allows evolution of any nonzero fractional population (expressed as a real number), no matter how small. The master-equation analysis thus serves as a model for evolutionary change in an infinite population. To provide an alternative approach as a control and to address effects of population size, we also performed stochastic Monte Carlo (MC) evolutionary simulations of finite, discrete populations (number of individuals are represented by integers).

Each MC simulation in the present effort tracks the evolution of $N = 1{,}000$ individuals. To facilitate comparison, the mutational parameters and other conditions for our MC simulations were chosen to closely resemble those in our master-equation analysis. Each MC simulation for a given network pair was initialized with all individuals carrying the same single prototype gene for the larger of the two neutral networks of the pair. Thus, for the network pair A and B, MC simulations were started with 1,000 copies of the single π_A gene. After initialization, successive rounds of mutations and selections were applied for 5,000 generations. In each generation, single-point mutations were randomly introduced at a probability of μ per residue by generating a random number $u \in [0, 1]$ for every H or P residue along every sequence in all 1,000 individuals (which can be either single-gene or two-gene genotypes). If $u \leq \mu$, an H → P or P → H mutation was made, depending on whether the original residue was H or P; otherwise the residue remained unchanged. We used $\mu = 0.001$ for our simulations. This stochastic mutation process produced a new set of 1,000 individuals, some or all of which could be mutated from the original sequences. Our definition of μ is identical to that in ref. 6, and is equivalent to μ_m/n, where μ_m is the total mutation rate in ref. 7.*

After each round of mutations, fitness was assigned to each of the 1,000 individuals in accordance with the fitness function described above. Sequences with $g > 6$ were assigned zero fitness. The next generation of 1,000 individuals were then selected by stochastically picking from the 1,000 individuals generated by the last round of mutation. In this selection process, the probability of an individual being picked was equal to its relative fitness. Specifically, the selection procedure was implemented as follows: The relative fitness of individual k in the population with fitness W_k is $W_k/\bar{W}$, where $\bar{W} = \sum_{k=1}^{N} W_k/N$. Let $R_0 \equiv 0$ and $R_k \equiv \sum_{k'=1}^{k} W_{k'}/N\bar{W}$ for $k = 1, 2, \ldots, N$; note that $R_N = 1$ by this definition. The R_k's are the boundaries of N discrete bins in $[0, 1]$ with sizes equal to the $W_k/N\bar{W}$'s. To select an individual, a random number $u \in [0, 1]$ was generated. Individual k was selected if and only if $R_{k-1} < u \le R_k$. This operation allows the random number to pick an individual by falling into one of the N bins. By repeating this operation $N = 1{,}000$ times, a new population of 1,000 individuals was selected. Because the same individual could be picked more than once by this procedure and some individuals might not be picked, fitter individuals would tend to be over-represented in the next generation.

Gene duplications were admitted after 100 generations. Subsequent to that time, a duplication attempt was made every 10 generations (or every 100 generations for the case of "low" duplication rate in Fig. S5*B*) by picking randomly one of the 1,000 individuals before the above-described selection procedure was applied. If the randomly picked individual was a single-gene genotype, it was turned into a two-gene genotype with a duplicated gene. This individual then carried two identical genes and could have altered fitness as a result. If the randomly picked individual was already a two-gene genotype, it remained unchanged. Irrespective of whether the duplication attempt resulted in a newly duplicated gene, the duplication attempt was repeated only after another 10 (or 100) generations. Because there are 1,000 individuals in the population, this stochastic procedure is equivalent to a duplication rate $\mu_d = (1/10) \times (1/1{,}000) = 10^{-4}$ (or $(1/100) \times (1/1{,}000) = 10^{-5}$).

Average properties from MC simulations were obtained from 100 independent runs (trajectories) simulated under identical conditions except different sets of random numbers were generated for the mutation, duplication, and selection steps. For each population of 1,000 individuals at a given time in our MC simulations, the frequencies of genotypes belonging to the following categories were recorded: (i) the prototype of the initial neutral network, (ii) other genes in the initial neutral network, (iii) single bridge genes, (iv) bridge pairs, and (v) neo/subfunctionalized pairs. The time-dependent average numbers of individuals in these genotype categories in the evolving populations were then determined by averaging over the 100 MC-simulated populations.

MC simulation results on the same two fitness landscapes considered in the main text are shown in Fig. S2*A*. Despite the differences between the master-equation and MC simulation procedures, results from the two approaches are qualitatively very similar, lending support for the robustness of our model predictions. The two approaches also produced similar results when two different duplication rates were compared (Fig. S5). In the SUBF scenario, adaptation is quick and mostly independent of gene duplications, whereas adaptation in the NEOF scenario is significantly impeded by a lower duplication rate.

An example MC simulation run, i.e., a single trajectory of SUBF (at the higher duplication rate of one per 10 generations) is provided in Fig. S2*B*. The genetic heterogeneity of the population changes over time (Fig. S2*B*, *Bottom*). It rose to the first peak before fixation of a single bridge gene at generation approximately 150 (cf. Fig. S2*A*, *Bottom*), fell sharply afterwards, and did not rise much again until just before the fixation of the duplicated bridge at around generation 500. This suggests that the population was initially spreading in different directions through the neutral network until a more beneficial gene (the bridge gene in this case) was reached and fixed. Gene duplication took longer to rise to high frequency because the fitness increase associated with duplication was not as high as that provided by fixation of the single bridge. In this example, subfunctionalization is a gradual process that increases genetic variation but does not lead to further significant adaptation (see the behavior after generation approximately 500, and especially after generation approximately 3,300 in Fig. S2*B*).

Generalizing to Other Target Structures and Neutral Networks. To evaluate the generality of our conclusions, network pairs other than the pair A and B were also used for simulations (see Figs. S3 and S6). Properties of all the networks used in the present study are listed in Table S1. Each neutral network encodes for a different HP protein structure. Because the density of states of a sequence and its reverse sequence are identical in the HP model, it is only necessary to consider neutral networks that cannot be obtained from one another by reversing all the sequences in the network. We chose the six largest neutral networks of $g = 1$ sequences in the 2D $n = 18$ HP model accordingly. Each of these networks was paired with a connected network sharing at least one bridge sequence with $g = 2$. As discussed above for network pair A and B, each network pair consists of all $g \le 6$ sequences that have either of the two target structures or both of the two target structures in their ground states.

To speed up simulations, a reduced set of genes obtained by removing all $g > 1$ genes that were not bridges in a given network was also considered. We used these reduced networks to simulate the results in Figs. S3 and S6. The removed genes were found to play no significant role during these simulations, as is evident from comparing the results computed for the full gene set in Figs. 2 and 3 of the main text and the corresponding results computed using the reduced gene set in Figs. S3*A* and S6*A* for the network pair A and B. The results computed for the reduced and full gene sets are virtually indistinguishable.

The simulation results we obtained are qualitatively very similar in all network pairs we investigated, with a few minor exceptions: Network pair I and J has a very stable bridge, which is transiently populated at a high frequency in its duplicated form even in the NEOF scenario (top plot in Fig. S3*E*). In contrast, network pair K and L are connected by bridges with particularly low stability that are not significantly populated before duplication during SUBF (Fig. S3*F*, *Bottom*).

For all the network pairs we studied, the steady-state populations of bridge gene pairs and subfunctionalized pairs after SUBF exhibited as two clearly distinct clusters in their scatter plots with the number of adjacent genotypes within Hamming distance 1 or Hamming distance 2 (Fig. S6). Regression curves for the subfunctionalized pairs are shown in Fig. S6. Dependence of steady-state population $(P_{ij})_{\mathrm{st}}$ on the number of sequence-space neighbors with Hamming distance 1 or Hamming distance 2 can be described approximately by a power law. The similarity among the six different network pairs in Fig. S6 of their separate clustering behavior of the bridge versus subfunctionalized pairs lends support to our conclusion in the main text that in general SUBF can

*Note that the statement "μ_m is equivalent to μ/n" in the second line below Eq. 1 on p. 812 of ref. 7 should read "μ_m is equivalent to $n\mu$". This error was merely typographical; it did not affect the results in ref. 7.

be a natural consequence of the mutational instability of the bridge genes.

A Control Study Using a Randomized Network Topology. As a control for the results shown in Fig. 3 in the main text, we devised a randomized network topology that has the same number of nodes (sequences) ω as in Fig. 1*A* in the main text but with randomized connections (edges) that are not based on an underlying biophysical chain model. The aim of using this control network was to assess the role of our biophysics-based model neutral network topology in the trend observed in the predicted distribution of steady state genotype populations, especially the conspicuous separation of bridge and subfunctionalized pairs in the scatter plots of $\ln(P_{ij})_{\mathrm{st}}$ versus number of adjacent sequence-space neighbors within Hamming distance 2 (see Fig. 3 in the main text and Fig. S6).

The average number of connections per node (i.e., its degree) $\langle A\rangle$ is 4.886 in the original topology. It follows that the probability of two nodes being adjacent would be $p_A = \langle A\rangle/(\omega - 1)$. Based on this statistics, a randomized adjacency matrix M of size $\omega \times \omega$ was created as follows. Because a sequence cannot be adjacent to itself, we set all diagonal elements of M to zero. For each off-diagonal element $m < n$ in the matrix, where $m, n = 1, 2, \ldots, \omega$ are row and column indices, a random number $u_{mn} \in [0, 1]$ was drawn from a uniform distribution. If $u_{mn} \le p_A$, nodes m and n was assigned to be adjacent by setting $M_{mn} = M_{nm} = 1$, otherwise we set $M_{mn} = M_{nm} = 0$. The resulting network had an actual average node degree of 4.859, which is essentially identical to the input $\langle A\rangle$ value of 4.886 as expected. Because Hamming distance as a sequence-space distance metric is not defined on such a randomized network, we defined another quantity that conceptually corresponds to the number of adjacent sequences within Hamming distance 2 from the randomized adjacency matrix M by considering M itself and its square, M^2. For a given node m, we took this quantity as the number of all nonzero off-diagonal elements $(M + M^2)_{mn}$ of the sum of matrices M and M^2. This quantity is the conceptual equivalent of the number of neighbors within Hamming distance 2 of node (sequence) m because a node $n \neq m$ with a nonzero element in M_{mn} is adjacent to m and a node $n \neq m$ with a nonzero element in $(M^2)_{mn}$ can be reached by node m in two steps in the randomized network defined by M.

When the randomized network was used to compute steady-state populations after SUBF, no separate clusters for the original bridge gene pairs and original subfunctionalized pairs were seen in the scatter plot of their logarithmic steady-state populations versus number of nodes reachable within two steps (see *Inset* of Fig. 3 in the main text). Moreover, the scatter in the data for the randomized network topology is so extensive that the approximate power-law dependence of steady-state population on the number of sequence-space neighbors within Hamming distance 2 observed for the result based on the HP model network topology is all but lost. These striking differences between the original and control calculations demonstrate clearly that the network topology of our model plays a central role in the relationship between mutational stability and evolutionary populations of bridge versus subfunctionalized pairs. This comparison also underscores the general importance of adopting network topologies that are underpinned by explicit-chain models based upon sound biophysical principles in the development of theory for molecular evolution.

The Dosage Parameter *d*. As described above and in the main text, the parameter d in our model is used to characterize the relationship between functional protein concentration, fractional protein population, and the number of copies of a given gene. Essentially, $d = 1$ corresponds to the assumption that there is a dosage increase (i.e., increase in protein concentration) upon gene duplication, whereas $d = 0$ corresponds to the assumption that there is no dosage increase (protein concentration remains the same) upon gene duplication. Negative dosage effects were not considered in the present study because in that case gene duplications are not likely to be retained.

The $d = 1$ case may be viewed as the default assumption for a duplicated gene that retains its promotor because the two gene copies can then be transcribed simultaneously to provide twice the amount of gene product. A situation related to, although not required for, the $d = 1$ case is that the concentration of the protein structure in question is suboptimal before gene duplication so that an increase in concentration is allowed by the cellular machinery. Suboptimal protein concentration before gene duplication could arise from trade-offs between multiple functional/structural states (26). Indeed, Kondrashov et al. have argued that some gene duplicates were retained because of a beneficial dosage increase (27, 28). In other words, the duplication itself is beneficial and thus will be preferentially retained. Similarly, Bergthorsson et al. (29) have also stated that the dosage increase of a suboptimal (promiscuous), but beneficial, enzyme function should render a gene duplication immediately advantageous. Although further research in this area is needed, positive dosage increase remains one of the most convincing explanations for duplicate retention, as its beneficial effect would take place immediately after the duplication and does not require additional mutations.

The $d = 0$ case corresponds to a situation in which the gene duplication itself is neutral. It may be viewed as a control to further elucidate the consequences of the $d = 1$ assumption. At the same time, our interest in the $d = 0$ case was also motivated by the argument of several authors that a gene duplication is of no intrinsic adaptive value in itself, and that their retention during NEOF or SUBF is purely by chance [see, e.g., reviews by Conant and Wolfe (30) and Innan and Kondrashov (31)]. With this in mind, the $d = 0$ case stipulates that the protein concentration remains unchanged upon the duplication of a gene i, i.e., $C_l(i, i) = \Phi(X_l, i)$ for $d = 0$. This requirement is satisfied by the general relationship in Eq. **S2**, viz., $C_l(i, j) = \max[\Phi(X_l, i), \Phi(X_l, j)]$, because this implies that $C_l(i, i) = \max[\Phi(X_l, i), \Phi(X_l, i)] = \Phi(X_l, i)$ at the duplication step. After the duplication event, the general $C_l(i, j)$ expression stipulates that only the sequence with a higher Φ contributes to functional concentration. As mentioned above, if the biological function of a protein structure is dependent upon its kinetic stability (23), it is reasonable to assign a higher functional concentration to a population with a higher kinetic stability than a population with a lower kinetic stability when the total residence times in the beneficial structure are identical in the two populations. Because the unfolding rate of a protein is often negatively correlated with the thermodynamic stability of its native state (32) and a higher unfolding rate means a lower kinetic stability, kinetic stability is expected to be positively correlated with thermodynamic stability. It follows that if the biological function of a protein is positively correlated with its kinetic stability, a population with a higher thermodynamic stability is expected to contribute more to the functional concentration than an equal population with a lower thermodynamic stability. In this context, our $d = 0$ assignment of functional concentration may be viewed as a drastic prescription that nonetheless embodies the above biophysical trend. A direct consequence of this preference for genes that encode thermodynamically more stable beneficial structures is that in general a higher fitness is assigned to SUBF pairs than to bridge pairs. Indeed, using this setup, our analysis demonstrated that a neutral gene duplication imposes a strong selection pressure for SUBF to occur rapidly because SUBF is the only way to increase fitness when $d = 0$ (Fig. 2*D* in the main text). This observation is of crucial relevance to the present study because adaptive SUBF is a feature of the EAC scenario (30, 31).

Biologically, the neutral duplication scenario embodied by our $d = 0$ case may apply in some situations in which protein concentrations after gene duplication can be maintained essentially at preduplication levels by regulatory mechanisms. In general, the homeostasis of certain proteins can be well-preserved by negative feedback loops (33), meaning that a higher concentration of a protein will cause its own down-regulation. This mechanism should apply after a gene duplication as well. An example of such a general process is dosage compensation in the silencing of one copy of the X chromosome in female human cells (34, 35). Another example is that extra gene copies are down-regulated by DNA methylation after gene duplications in mammals (36). A reduced expression of duplicates was also observed due to mutations in upstream regions of genes (37). Evidence for dosage balancing mechanisms was also found in yeast, wherein only 15% of genes are detrimental when overexpressed by increased gene copy numbers on plasmids. The majority of genes do not produce a different phenotype when overexpressed. Based on this data, the authors of the study assume that gene regulatory feedback controls protein levels (38). Finally, whole genome duplications also provide conditions that render the overall relative stoichiometry of biomolecules constant, thus resulting in an effectively neutral gene duplication (39–41). The $d = 0$ case may serve to capture the lack of dosage effect at the duplication step in situations similar to these, although gene regulation is not explicitly modeled in our analysis.

It is important, however, to emphasize that the $d = 0$ and $d = 1$ cases represent extreme situations that serve to bracket a range of possible situations. In reality, even when a gene duplication increases dosage, it may not exactly double the level of protein as assumed for the $d = 1$ case, because both genes may need to use the same transcriptional activators that may in turn be limited in concentration so that expression increase may be less than twofold. Indeed, the limited concentration of transcription factors that presumably were evolved to be fine tuned for a single gene locus may explain the near absence of a dosage increase upon gene duplication in some situations, which would correspond more closely to the $d = 0$ case in this respect. As discussed above, the present $d = 0$ and $d = 1$ formulations also embody different relationships between functional concentration and protein population. In reality, one expects the degree to which functional concentration depends on thermodynamic stability and the resulting kinetic stability to vary from protein to protein, depending on the biological function in question. In view of these considerations, future work will need to extend the treatment of dosage effect from the present binary choices to include a continuum of possibilities between extremes exemplified by the $d = 0$ and $d = 1$ cases here.

Experimental Evidence for Adaptation Before and After Gene Duplication. It remains a challenge in many instances to match biological data to theoretical evolutionary scenarios such as DDC versus EAC, or NEOF versus EAC (30). Ratios of nonsynonymous over synonymous rates of nucleotide substitutions, K_a/K_s (also denoted as d_N/d_S or ω) (42) have been used to study the evolution of gene duplicates. (The ω here should not be confused with the variable for the number of genes in the above discussion.) Although such methods are likely to be too crude to distinguish between different types of SUBF, some general patterns of K_a/K_s are useful in distinguishing SUBF from NEOF.

For example, a K_a/K_s study has shown that in many cases of successful gene duplications both paralogs have $K_a/K_s < 1$, implying that they are under relaxed purifying selection, i.e., they tolerate more mutations (27). In general, a K_a/K_s value between 0 and 1 would correspond to neutral evolution constrained within a neutral network with some nonsynonymous substitutions being tolerated. An example of EAC is provided by Des Marais and Rausher for the dihydroflavonol-4-reductase (DFR) genes in plants, which are part of the anthocyanin pathway (43). The authors provided a clear criteria for identifying EAC and distinguishing it from NEOF: First, both duplicates have to evolve equally by adaptive changes during EAC, whereas in NEOF only one copy evolves. Second, an ancestral function is improved after gene duplication in EAC, whereas it is not in NEOF. They also measured the evolutionary rates along the phylogeny of DFR genes. In one branch leading towards a gene duplication (resulting in paralogs DFR-A and DFR-C), $K_a/K_s \gg 1$ was found, indicating that the ancestor was undergoing adaptive evolution. This finding is consistent with adaptive conflict, wherein single amino acid substitutions improve a new function at the expense of the old function. After duplication, purifying selection was observed in the two duplicates, leading to divergent and an improvement of ancestral functions (43).

During the course of the present investigation, we found a potential case of EAC for which functional data of a multifunctional ancestor and DNA sequence alignments are available in the literature (44, 45). The proteins in question are a family of fluorescent proteins in corals. Matz and colleagues reconstructed the evolution of these proteins. Without identifying their finding with any evolutionary model, they experimentally reconstructed the ancestor of a gene duplication leading towards green color in one copy and red color in the other copy. The ancestor was found to exhibit a dual phenotype: A fraction of proteins emitted green light while another fraction emitted red light (44, 45). Although it is not certain whether such an ancestral protein existed in natural corals, it is remarkable that it was found in exactly the position of the phylogeny where it would be expected by EAC. Seven of the residues that were changed during color adaptation are thought to exert their color-changing effect through changes of the overall protein fold (46), which fits the structure-function assumption adopted in our model.

We have performed a K_a/K_s analysis of the phylogeny in Ugalde et al. (45) based on the DNA alignment of Kelmanson et al. (44) by using a method of Kosakovsky Pond and Frost (47). The latter method assigns K_a/K_s to every branch in the phylogeny. In particular, we found that the branch leading to the multifunctional red/green ancestor shows strong signs of positive selection ($K_a/K_s \gg 1$). The subsequent branch leading to the pre-red ancestor also has $K_a/K_s \gg 1$ before duplication, which then is reduced to $K_a/K_s \approx 0.8$ for both the green and the red branches after duplication. These results match the analysis of Matz and colleagues (48). The duplicates of the pre-red ancestor specialized on different wavelengths (green and red), indicating that the initial adaptive conflict was somewhat resolved. This behavior is consistent with EAC.

More recently, EAC was also indicated in the phylogeny of plant enzymes that generate secondary metabolites by constructing several ancestors and experimental measurements of enzymatic activities (49). Positive evolution was found in ancestral branches before gene duplication, which is indicative of a shift in selection pressures that might have led to adaptive conflict. Consistent with the EAC scenario, ancestral nonpreferred enzymatic activities before gene duplication were enhanced in one of the daughter enzymes after duplication (49).

Population Size, Quasi-Species, and Evolutionary Time Scale. As discussed above, the effective population size in our master-equation approach (Eqs. **S4** and **S5**) is an infinite number of individuals. Therefore, the process modeled by the master-equation approach corresponds to a quasi-species regime of evolution (50–52) in which mutations are frequent. Quasi-species are found when the product of total mutation rate μ_m and population size N (i.e., the expected number of mutations per generation) is very large ($\mu_m N \gg 1$). Under these conditions, many co-existing gene variants arise, instead of just a single predominant variant of a gene within a population when $\mu_m N \ll 1$. The evolutionary system described in Eqs. **S4** and **S5** thus resembles fast-replicating, asexual organisms with large population sizes, such as viruses and bacteria.

A small $\mu_m N$ is not a fundamental impediment against adaptation that can proceed one mutation at a time, except that evolution will be slower and that alternative mechanisms such as recombination might be more dominant. However, an epistatic barrier such as that seen in Fig. 2*F* would become a formidable obstacle if $\mu_m N \ll 1$. In that case, SUBF is less likely to occur by neutral drift alone but would require some adaptive pressure. In this regard, phenotypic mutations (53–55) might be an additional mechanism to overcome this barrier.

A case of small $\mu_m N$ (but not $\mu_m N \ll 1$) is provided by our MC simulations. Because there are $n = 18$ residues per sequence in our 2D HP model and there are $N = 1{,}000$ individuals in the population, for $\mu = 0.001$, the number of mutations per generation is between $\mu_m N = n\mu N = 18$ (when all individuals have a single gene) and 36 (when all individuals have two genes). Moreover, many of the mutations in our MC simulation were lethal because the ground states of the resulting sequences contain neither of the two target structures. Consequently, the number of genotypes populated at any time in our MC simulations is very limited compared to that in the effectively $N \to \infty$ "quasi-species" population described by our master-equation formulation.

Consistent with the above considerations, we found that evolution proceeds at a significantly slower pace in our MC than in our master-equation simulations. As discussed in the main text, evolutionary processes predicted by the two approaches for the same network pair using the same mutation and duplication rates and identical fitness functions exhibit very similar qualitative features. This similarity is readily seen by comparing, e.g., the master-equation results in Fig. 2 *C* and *D* in the main text with the MC results in Fig. S2*A*. Also clear from this comparison, however, is that the rate of evolution is much slower in the MC than in the master-equation simulations. For instance, the population peak for the bridge pair in the SUBF case appears at generation approximately 80 in the master-equation approach (Fig. 2*D* in main text) but the corresponding peak appears at generation approximately 600 in the MC simulation (Fig. S2*A*, lower plot). The relative slowdown in the MC simulation is significant even taking into account the absence of duplication in the first 100 generations. To gain further understanding, we have also performed a MC simulation under the same conditions as those for the lower plot in Fig. S2*A* except that the number of trajectories used for averaging was 30 (instead of 100, for computational efficiency) and duplication was attempted at every generation in a randomly chosen individual at a probability of 0.1 (thus corresponding to a duplication rate of $\mu_d = 10^{-4}$), instead of attempting duplication regularly every 10 generations after the first 100 generations in a randomly chosen individual. This MC simulation had no time lag in duplication attempts and duplication was attempted in every generation such that the dynamic conditions of the MC simulation are essentially identical to those of the master-equation formulation for Fig. 2*D* of the main text. Even so, the population peak for the bridge pair appears at generation approximately 400 in this case, still significantly later than the approximately 80 generations needed to achieve the corresponding peak in the master-equation formulation. Taken together, our observations indicate that population size has little effect on the general relationships elucidated here between selection pressure and dosage effect on one hand and the NEOF and SUBF scenarios on the other; but the effectively infinite population in the master-equation formulation allows for much speedier evolutionary changes than the finite populations in our MC simulations.

1. Lau KF, Dill KA (1989) A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* 22:3986–3997.
2. Dill KA (1990) Dominant forces in protein folding. *Biochemistry* 29:7133–7155.
3. Lau KF, Dill KA (1990) Theory for protein mutability and biogenesis. *Proc Natl Acad Sci USA* 87:638–642.
4. Chan HS, Dill KA (1991) Sequence space soup of proteins and copolymers. *J Chem Phys* 95:3775–3787.
5. Dill KA, et al. (1995) Principles of protein folding—a perspective from simple exact models. *Prot Sci* 4:561–602.
6. Bornberg-Bauer E, Chan HS (1999) Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. *Proc Natl Acad Sci USA* 96:10689–10694.
7. Cui Y, Wong WH, Bornberg-Bauer E, Chan HS (2002) Recombinatoric exploration of novel folded structures: A heteropolymer-based model of protein evolutionary landscapes. *Proc Natl Acad Sci USA* 99:809–814.
8. Wroe R, Bornberg-Bauer E, Chan HS (2005) Comparing folding codes in simple heteropolymer models of protein evolutionary landscape: Robustness of the superfunnel paradigm. *Biophys J* 88:118–131.
9. Wroe R, Chan HS, Bornberg-Bauer E (2007) A structural model of latent evolutionary potentials underlying neutral networks in proteins. *HFSP J* 1:79–87.
10. Chen T, Vernazobres D, Yomo T, Bornberg-Bauer E, Chan HS (2010) Evolvability and single-genotype fluctuation in phenotypic properties: A simple heteropolymer model. *Biophys J* 98:2487–2496.
11. Irbäck A, Peterson C, Potthast F (1996) Evidence for nonrandom hydrophobicity structures in protein chains. *Proc Natl Acad Sci USA* 93:9533–9538.
12. Irbäck A, Sandelin E (2000) On hydrophobicity correlations in protein chains. *Biophys J* 79:2252–2258.
13. Chan HS, Bornberg-Bauer E (2002) Perspectives on protein evolution from simple exact models. *Appl Bioinformatics* 1:121–144.
14. Chan HS, Kaya H, Shimizu S (2002) in *Current Topics in Computational Molecular Biology*, eds Jiang T, Xu Y, Zhang MQ (MIT Press, Cambridge, MA), pp 403–447.
15. Chan HS (2000) Modeling protein density of states: Additive hydrophobic effects are insufficient for calorimetric two-state cooperativity. *Proteins* 40:543–571.
16. Chan HS, Zhang Z, Wallin S, Liu Z (2011) Cooperativity, local-nonlocal coupling, and nonnative interactions: Principles of protein folding from coarse-grained models. *Annu Rev Phys Chem* 62:301–326.
17. Govindarajan S, Goldstein RA (1997) Evolution of model proteins on a foldability landscape. *Proteins* 29:461–466.
18. Hirst JD (1999) The evolutionary landscape of functional model proteins. *Protein Eng.* 12:721–726.
19. Xia Y, Levitt M (2002) Roles of mutation and recombination in the evolution of protein thermodynamics. *Proc Natl Acad Sci USA* 99:10382–10387.
20. Bloom JD, Wilke CO, Arnold FH, Adami C (2004) Stability and the evolvability of function in a model protein. *Biophys J* 86:2758–2764.
21. Xia Y, Levitt M (2004) Simulating protein evolution in sequence and structure space. *Curr Opin Struct Biol* 14:202–207.
22. Bornberg-Bauer E (1997) How are model protein structures distributed in sequence space? *Biophys J* 73:2393–2403.
23. Sanchez-Ruiz JM (2010) Protein kinetic stability. *Biophys Chem* 148:1–15.
24. Peña MI, Van Itallie E, Bennett MR, Shamoo Y (2010) Evolution of a single gene highlights the complexity underlying molecular descriptions of fitness. *Chaos* 20:026107.
25. Fruchterman TMJ, Reingold EM (1991) Graph drawing by force-directed placement. *Software Pract Exper* 21:1129–1164.
26. Khersonsky O, Tawfik DS (2010) Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annu Rev Biochem* 79:471–505.
27. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. *Genome Biol* 3:RESEARCH0008.
28. Kondrashov FA, Kondrashov AS (2006) Role of selection in fixation of gene duplications. *J Theor Biol* 239:141–151.
29. Bergthorsson U, Andersson DI, Roth JR (2007) Ohno's dilemma: Evolution of new genes under continuous selection. *Proc Natl Acad Sci USA* 104:17004–17009.
30. Conant GC, Wolfe KH (2008) Turning a hobby into a job: How duplicated genes find new functions. *Nat Rev Genet* 9:938–950.
31. Innan H, Kondrashov FA (2010) The evolution of gene duplications: Classifying and distinguishing between models. *Nat Rev Genet* 11:97–108.
32. Matthews CR, Hurle MR (1987) Mutant sequences as probes of protein folding mechanisms. *BioEssays* 6:254–257.
33. Acar M, Pando BF, Arnold FH, Elowitz MB, van Oudenaarden A (2010) A general mechanism for network-dosage compensation in gene circuits. *Science* 329:1656–1660.
34. Birchler JA, Bhadra U, Bhadra MP, Auger DL (2001) Dosage-dependent gene regulation in multicellular eukaryotes: Implications for dosage compensation, aneuploid syndromes, and quantitative traits. *Dev Biol* 234:275–288.
35. Straub T, Becker PB (2007) Dosage compensation: The beginning and end of generalization. *Nat Rev Genet* 8:47–57.
36. Chang AY-F, Liao B-Y (2011) DNA methylation rebalances gene dosage after mammalian gene duplications. *Mol Biol Evol* 29:133–144.
37. Qian W, Liao B-Y, Chang AY-F, Zhang J (2010) Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet* 26:425–430.
38. Sopko R et al. (2006) Mapping pathways and phenotypes by systematic gene overexpression. *Mol Cell* 21:319–330.
39. Veitia RA (2005) Paralogs in polyploids: One for all and all for one? *Plant Cell* 17:4–11.
40. Otto SP (2007) The evolutionary consequences of polyploidy. *Cell* 131:452–462.
41. Sémon M, Wolfe KH (2007) Consequences of genome duplication. *Curr Opin Genet Dev* 17:505–512.
42. Hurst LD (2002) The K_a/K_s ratio: Diagnosing the form of sequence evolution. *Trends Genet* 18:486–487.
43. Des Marais DL, Rausher MD (2008) Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* 454:762–765.

44. Kelmanson IV, Matz MV (2003) Molecular basis and evolutionary origins of color diversity in great star coral Montastraea cavernosa (Scleractinia: Faviida). *Mol Biol Evol* 20:1125–1133.
45. Ugalde JA, Chang BSW, Matz MV (2004) Evolution of coral pigments recreated. *Science* 305:1433.
46. Field SF, Matz MV (2010) Retracing evolution of red fluorescence in GFP-like proteins from Faviina corals. *Mol Biol Evol* 27:225–233.
47. Kosakovsky Pond SL, Frost SDW (2005) A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Mol Biol Evol* 22:478–485.
48. Field SF, Bulina MY, Kelmanson IV, Bielawski JP, Matz MV (2006) Adaptive evolution of multicolored fluorescent proteins in reef-building corals. *J Mol Evol* 62:332–339.
49. Huang R et al. (2012) Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates. *Proc Natl Acad Sci USA* 109:2966–2971.
50. Eigen M, McCaskill J, Schuster P (1988) Molecular quasi-species. *J Phys Chem* 92:6881–6891.
51. Wilke CO, Wang JL, Ofria C, Lenski RE, Adami C (2001) Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature* 412:331–333.
52. Elena SF, Carrasco P, Daròs JA, Sanjuán R (2006) Mechanisms of genetic robustness in RNA viruses. *EMBO Rep* 7:168–173.
53. Bürger R, Willensdorfer M, Nowak MA (2006) Why are phenotypic mutation rates much higher than genotypic mutation rates? *Genetics* 172:197–206.
54. Whitehead DJ, Wilke CO, Vernazobres D, Bornberg-Bauer E (2008) The look-ahead effect of phenotypic mutations. *Biol Direct* 3:18.
55. Goldsmith M, Tawfik DS (2009) Potential role of phenotypic mutations in the evolution of protein expression and stability. *Proc Natl Acad Sci USA* 106:6197–6202.

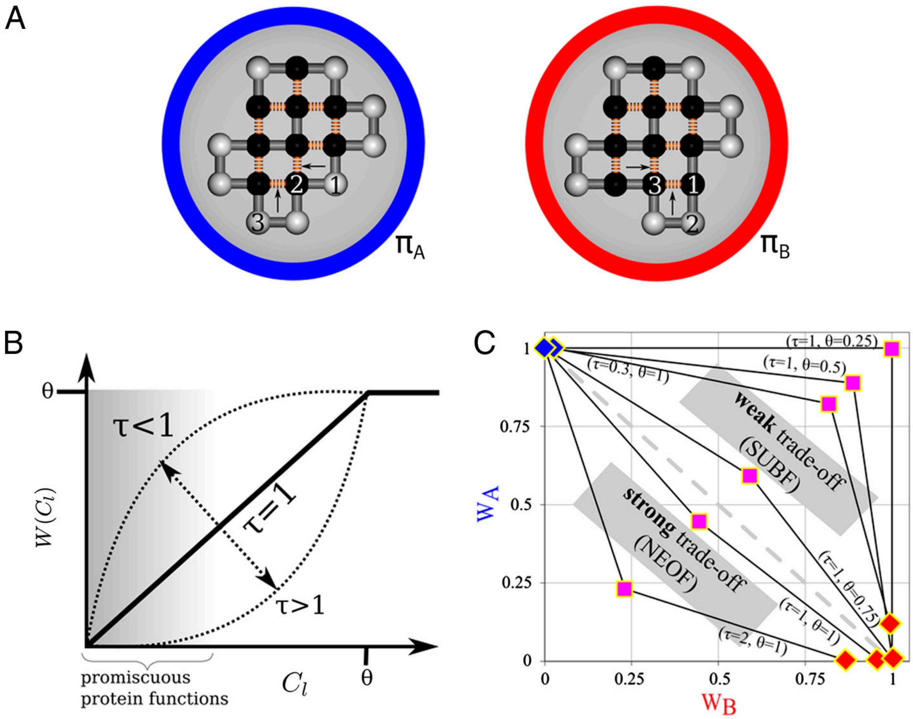


Fig. S1. Biophysical model of protein evolution. (*A*) Prototype sequences π_A and π_B of neutral networks A and B are shown in their respective native state conformations X_A and X_B. The protein chains are configured on a two-dimensional square lattice and the chains consist of hydrophobic (H) and polar (P) residues that are depicted, respectively, by dark and light beads. Hydrophobic-hydrophobic contacts are indicated by dashed orange connections. The two model proteins differ by three substitutions in their sequences (1, 2, and 3) and by two contacts in their structures (arrows). (*B*) Fitness W as a function of structural stability. The fitness contribution $W(C_l)$ of a beneficial protein structure X_l is modeled in this work as a function of its concentration C_l, which is in turn a function of the stability (fractional population) of X_l. As described in the *SI Text*, C_l is dependent upon the number of genes (one or two) in a given genotype; θ is an upper bound on W, i.e., fitness does not increase further with increasing C_l for $C_l > \theta$ in our model; and τ controls the deviation of $W(C_l)$ from a linear relationship in the interval $[0, \theta]$. Promiscuous protein functions are associated with excited-state structures at low concentrations (shaded region). (*C*) Fitness trade-offs between adopting structures X_A and X_B in Fig. 2*A* of the main text under various sets of (τ, θ) parameter values. W_A and W_B are fitness contributions from X_A and X_B, respectively, and are plotted here in units of θ (i.e., to facilitate comparison, fitness values for different (τ, θ) sets are normalized so that $W_A(\pi_A) = 1$ in all cases; exact fitness values for all data points in this plot are provided in Table S1). Fitness values are shown for three genes: prototype π_A (blue diamond), bridge β_{AB} (magenta square), and prototype π_B (red diamond). Connecting lines are used to indicate data points for the same (τ, θ) set of parameter values. If the total (combined) fitness $W_A + W_B$ is higher for the generalist sequence β_{AB} than that for either of the specialist sequences π_A and π_B, the trade-off between X_A and X_B is weak. Otherwise the trade-off is considered to be strong. The two trade-off regimes are demarcated by the dashed line.

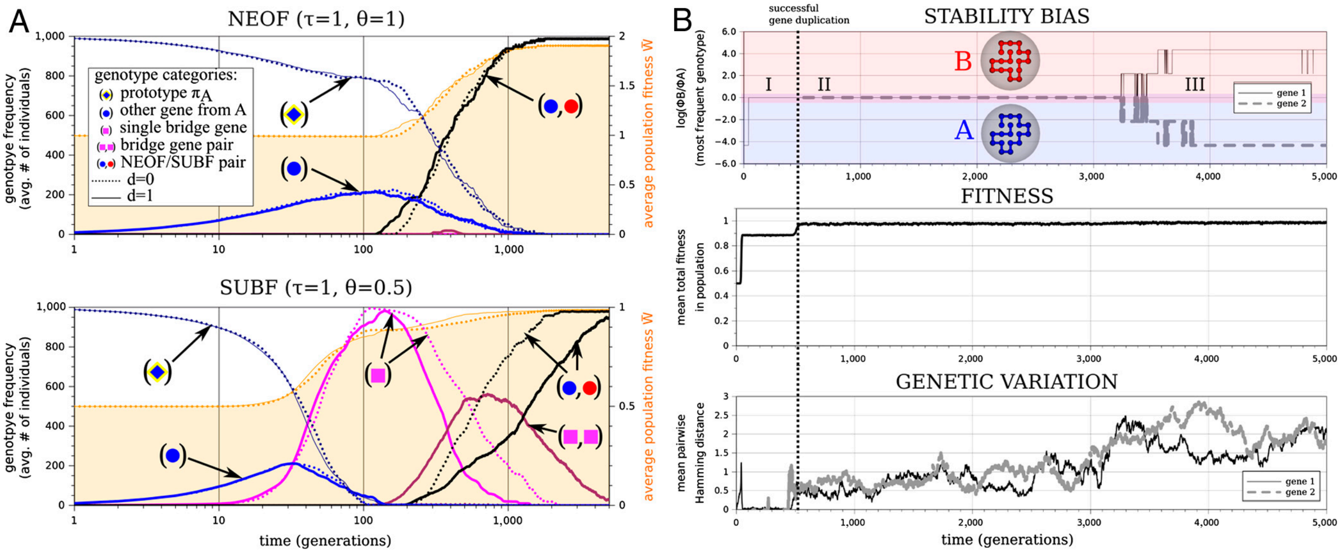


Fig. S2. Monte Carlo (MC) simulations show NEOF and SUBF in finite evolving populations. As described in *SI Text*, each simulation run involves 1,000 individuals. Every individual was initialized (at generation 0) as a single-gene genotype carrying a copy of π_A. These sequences were allowed to evolve stochastically in subsequent generations under a fitness function that selected for both the target structures X_A and X_B (shown in Fig. 1*A* and Fig. S1*A*). After 100 generations, gene duplication was imposed on a randomly chosen individual every 10 generations as long as single-gene genotypes (loci) existed among the 1,000 individuals in the simulation. After a duplication event, a second gene locus was occupied in an individual and this second sequence was then allowed to further mutate. (*A*) Simulations were performed using the same sets of τ, θ fitness parameters as those in Fig. 2 in the main text. Results were averaged over 100 independent runs (i.e., 100 independently evolving populations of 1,000 individuals each). As for the master-equation results in Fig. 2 of the main text, MC simulations were performed for the case with dosage effect ($d = 1$; solid curves) and also for the case without dosage effect ($d = 0$; dotted curves). Populations of various genotypes are plotted in different colors (as indicated) using the left vertical scale; average fitness values are plotted in orange according to the right vertical scale. *Top* NEOF is a consequence of strong selection pressure. Population fitness is seen to increase only upon duplication and divergence. *Bottom* A lower selection pressure allows SUBF via multifunctional bridge intermediates. Population fitness increases early with the rise of transient populations of single and duplicated bridge sequences. The fitness value averaged over generations 2,000 to 2,100 and over generations 4,000 to 4,100 are 0.9827 and 0.9844, respectively. (*B*) Analysis of an example MC simulation showing SUBF. Data are presented for a single MC run using $\tau = 1$, $\theta = 0.5$, and $d = 1$ under the same conditions as those in *A* (i.e., results here are for one of the 100 runs considered in the Bottom of *A*). Initially only one gene locus was populated (gene 1; black solid line), a second gene could then arise by duplication (gene 2; gray dashed lines). Three properties of the evolving population of 1,000 individuals are shown as functions of the number of generations. *Top* The stability bias $\log_{10}(\Phi_B/\Phi_A)$ of the most frequent genotype. The onset of a zero bias indicates that a bistable bridge protein has evolved from the initial π_A (phase I). The commencement of the gray dashed line indicates when the duplication of a bridge gene was fixed, i.e., became the most frequent genotype in the population (dotted vertical line; start of phase II). Eventually, the divergence of the solid and dashed lines signals that SUBF was fixed (phase III). *Middle* The mean population fitness increases significantly twice during SUBF, indicating that it is an adaptive process in this situation. Fitness is near optimal in the plateau region after the second significant increase but it still increases very gradually. For instance, the fitness values averaged over generations 2,000 to 2,100 and over generations 4,000 to 4,100 are 0.978 and 0.984, respectively. *Bottom* Genetic variation is measured as the average pairwise Hamming distance. This property was computed separately for sequences in gene 1 and for sequences in gene 2. Significant genetic variations among sequences in gene 2 are seen both before and after the fixation of the duplication of a bridge gene.

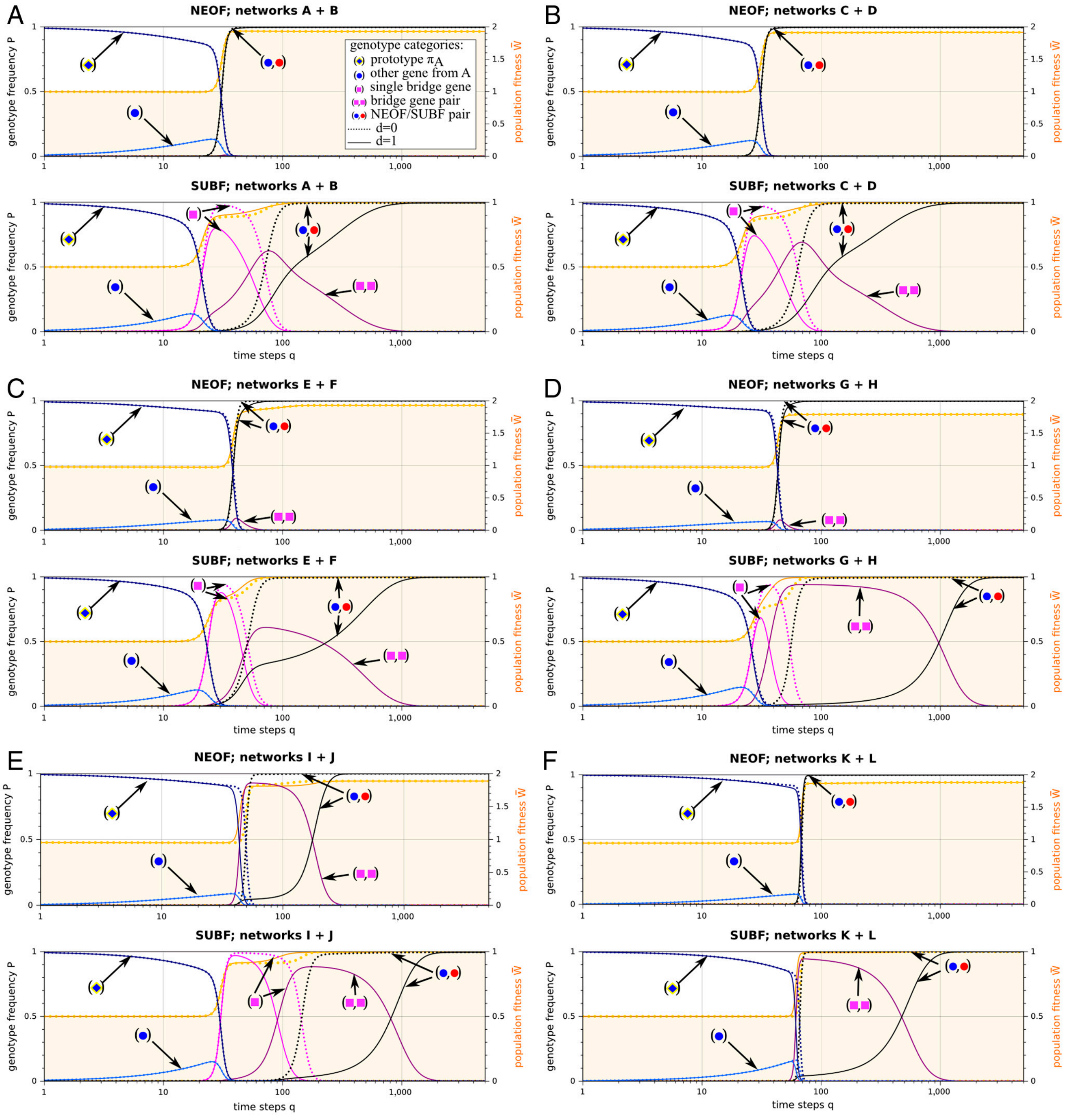


Fig. S3. Generality of the conditions for NEOF and SUBF. The trend seen in Fig. 2 of the main text for networks A and B is verified for other networks in our model. Shown in this figure are results from the master equation method (Eqs. **S4** and **S5**). For *A*–*F*, $\tau = 1$, $\theta = 1$, and $\tau = 1$, $\theta = 0.5$ were used for the upper and lower plots, respectively. Results were computed using reduced gene sets as described in the *SI Text*. The quantities plotted are equivalent to, and are plotted in the same style as those in Fig. 2 *C* and *D* in the main text. Corresponding results for the other network pairs in Table S1 are shown in *B*–*F*, respectively, for C and D, E and F, G and H, I and J, and K and L.

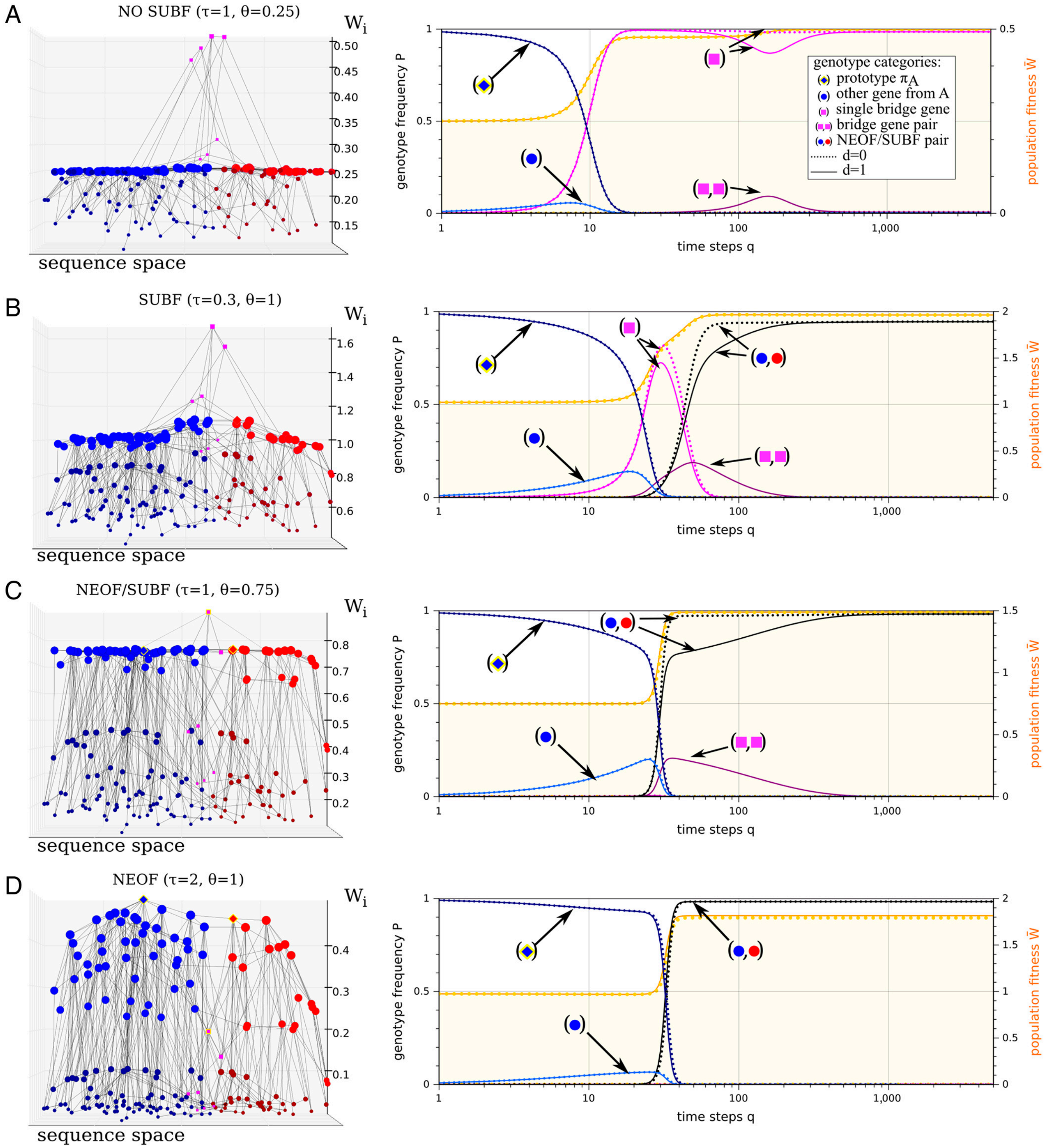


Fig. S4. Additional fitness landscapes and simulations with different parameter sets. Shown here are four additional examples of weak and strong trade-offs and the resulting evolutionary dynamics (parameter sets from Fig. S1*C*). Except for the different combinations of τ, θ parameters, results here were obtained using the same master-equation formulation for the network pair A and B and are presented in the same style as in Fig. 2 of the main text. (*A*) A low selection pressure ($\tau = 1$, $\theta = 0.25$) leads to the permanent retention of a single bridge gene as the dominant genotype. Duplications of bridges are only transiently successful. (*B*) Under a fitness function that favors promiscuous functions ($\tau = 0.3$, $\theta = 1$; cf. the $\tau < 1$ case in Fig. S1*B*), excited structural states have a stronger impact on fitness even when their populations are relatively small. Because multifunctionality is particularly rewarded, SUBF can proceed rapidly regardless of whether dosage effect is present ($d = 1$) or absent ($d = 0$). (*C*) A fitness landscape where a bridge is only slightly fitter than nonbridges (slightly weak trade-off) leads to an evolutionary process that bears features of both NEOF (single bridge not populated) and SUBF (bridge pairs populated when $d = 1$). (*D*) NEOF arises also when $\tau = 2$, which leads to a stronger trade-off (see Fig. S1*C*).

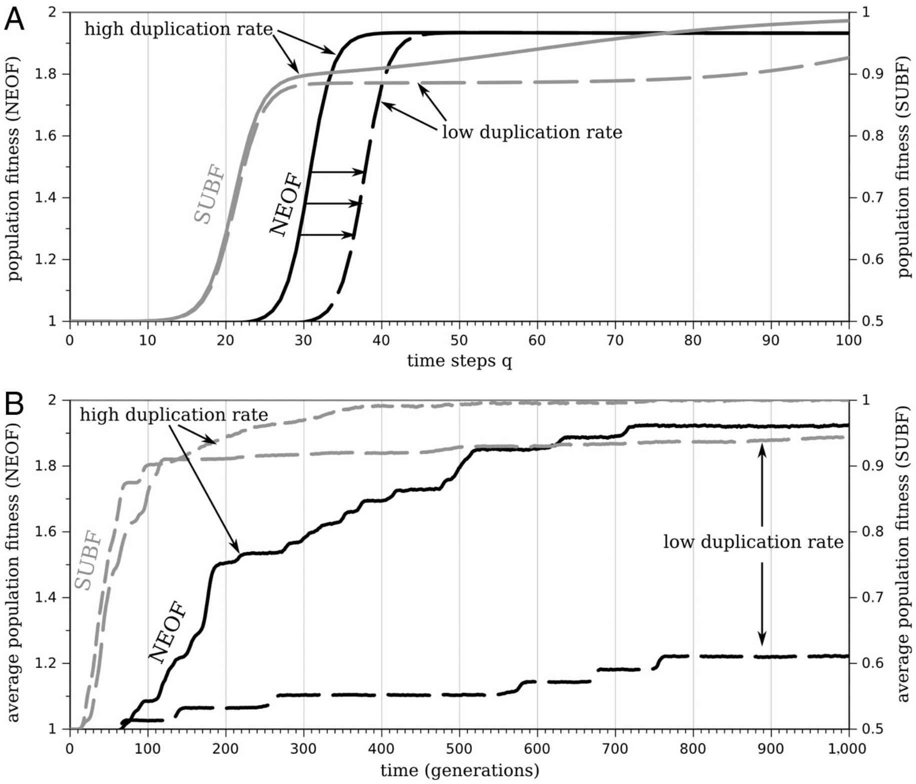


Fig. S5. Multifunctionality allows for rapid adaptation irrespective of gene duplication rate, but is particularly advantageous when gene duplications are rare. Using a mutation rate of $\mu = 10^{-3}$ throughout, evolutionary data from our master-equation formulation (*A*) and MC simulation (*B*) in the presence of dosage effect ($d = 1$) were obtained for the network pair A and B for the NEOF ($\tau = 1$, $\theta = 1$, black curves; fitness given by the left vertical scale) and SUBF ($\tau = 1$, $\theta = 0.5$, gray curves; fitness given by the right vertical scale) scenarios under a duplication rate that is either relatively high (solid curves) or relatively low (dashed curves). (*A*) Master-equation results for infinite population. High and low duplication rates here correspond, respectively, to $\mu_d = 10^{-4}$ and $\mu_d = 10^{-6}$. Adaptation occurs earlier in SUBF than in NEOF. Multifunctionality can increase before gene duplication in SUBF. Thus, for SUBF, the time needed for adaptation to commence is not lengthened by a lower duplication rate. Adaptation is indicated by a rapid initial increase in normalized average population fitness $\bar{W}$. After achieving a high level of fitness, further increase in fitness is sensitive to duplication rate because this second-stage increase is a result of the positive dosage effect afforded by duplicated bridge genes. By comparison, adaptation in NEOF is a more lengthy process because in this case adaptation takes place only after gene duplication. As a result, the process is slowed down significantly by the decrease in duplication rate (underscored by the horizontal arrows). In other words, the time to achieve adaptation (high fitness) increases with a lower duplication rate in general. However, this delay is much more prominent for NEOF than for SUBF because duplication is not a prerequisite for adaptation in SUBF when a single multifunctional gene can provide two functions at sufficient levels. In contrast to the adaptation process in SUBF, adaptation in NEOF reaches the maximum fitness rapidly in a sigmoidal manner soon after average population fitness begins to increase. (*B*) MC simulation results for evolving populations of 1,000 individuals. The average population fitness were averaged from 100 independent simulation runs. Low and high duplication rates correspond, respectively, to one duplication every 10 and 100 generations. The general trend exhibited is consistent with that in *A*. For the small populations analyzed here, however, SUBF is seen to be much more adaptive than NEOF when gene duplications are rare.

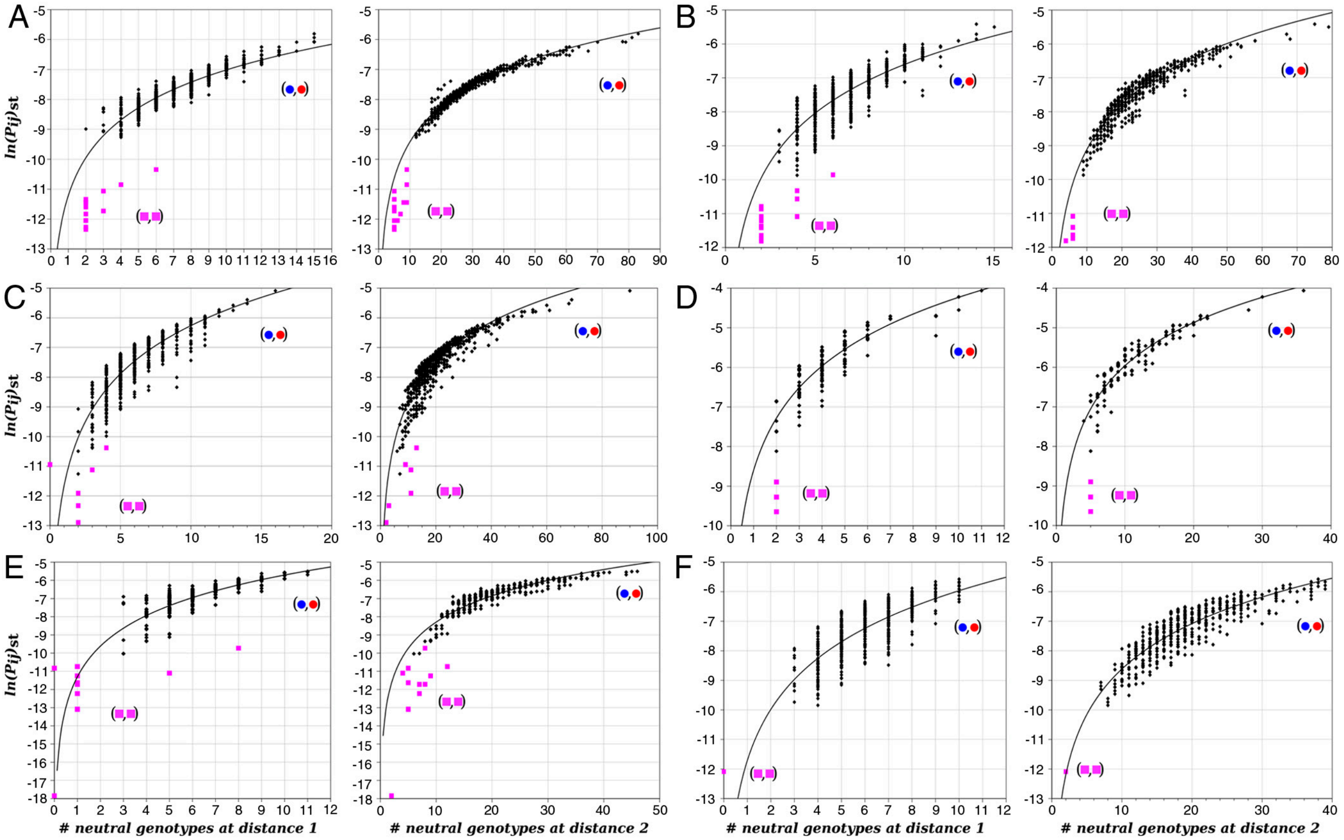


Fig. S6. Mutationally unstable bridge sequences are sparsely populated in SUBF. Steady state populations $(P_{ij})_{\mathrm{st}}$ of genotypes with maximum fitness $W_{ij} = 1$ in the SUBF scenario were computed using $\tau = 1$, $\theta = 0.5$, and $d = 1$ as in Fig. 3, except we now broaden our consideration to six different network pairs and reduced gene sets were used for their computational efficiency (see *SI Text*) to obtain the results in this figure. For a given network pair in *A*–*F*, the scatter plots of $\ln(P_{ij})_{\mathrm{st}}$ versus the number of genotypes within Hamming distance 1 and within Hamming distance 2 from (i, j) are shown, respectively, on the left and right of the panel. As in Fig. 3, bridge and subfunctionalized gene pairs are shown, respectively, as magenta squares and black diamonds. Data points for the subfunctionalized gene pairs were fitted to $y = m \ln x + b$, where y denotes $\ln(P_{ij})_{\mathrm{st}}$ and x represents the number of neutral genotypes within Hamming distance 1 or 2 from (i, j) in the given network pair. The corresponding least-squares fits are shown as continuous curves in the scatter plots. Results were obtained for all six combinations of neutral networks in Table S1. The corresponding numbers of subfunctionalized pairs and bridge pairs are listed, respectively, as the first and second entries in parentheses as follows: (*A*) A and B (1,728 and 24); (*B*) C and D (1,480 and 28); (*C*) E and F (1,343 and 10); (*D*) G and H (276 and 4); (*E*) I and J (802 and 14); (*F*) K and L (1,092 and 1). For every network pair we considered, the scatter plot of $\ln(P_{ij})_{\mathrm{st}}$ versus Hamming distance 2 consistently show all data points clustering quite tightly around the fitted curve. This observation suggests that the bridge and subfunctionalized pairs are governed by the same approximate power-law relationship between $(P_{ij})_{\mathrm{st}}$ and the number of neutral neighbors within Hamming distance 2. In this perspective, the low $(P_{ij})_{\mathrm{st}}$ values (thus low $\ln(P_{ij})_{\mathrm{st}}$ values) for the bridge pairs are seen as a natural consequence of their low mutational robustness.

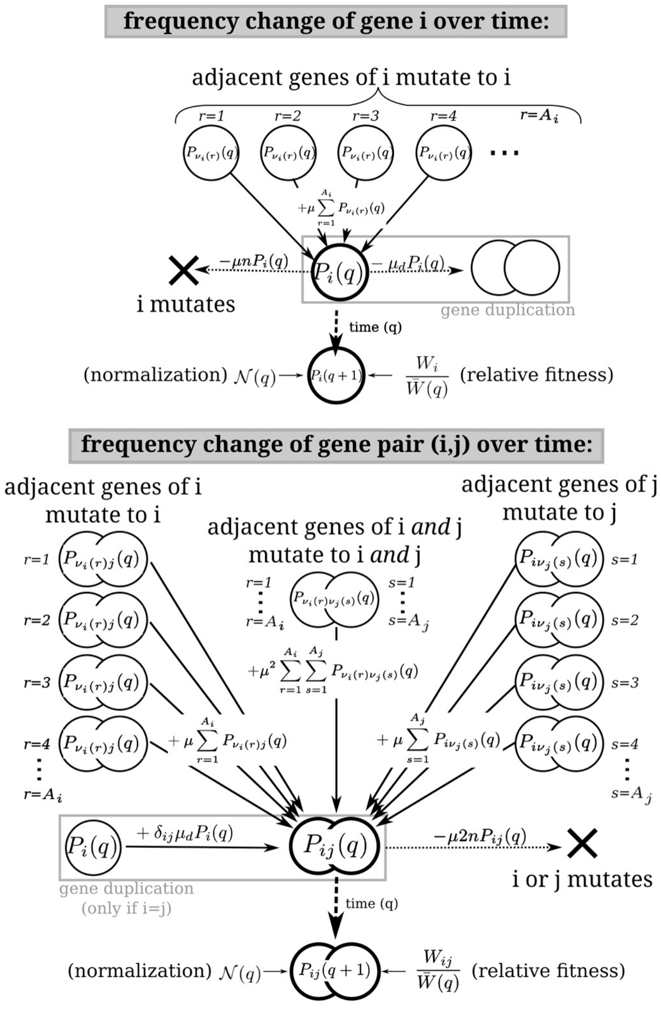
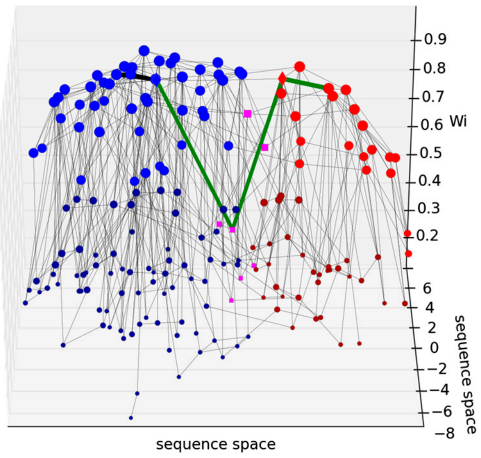
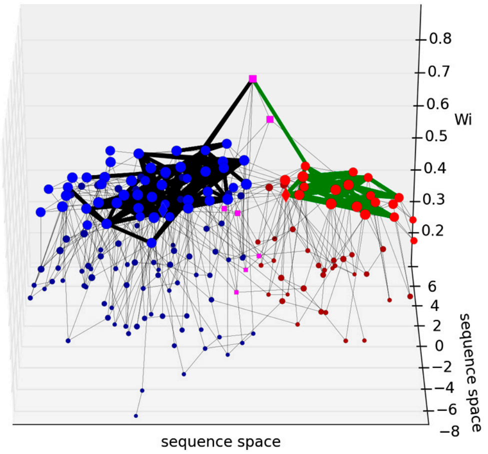

frequency change of gene i over time:
adjacent genes of i mutate to i
r=1
r=2
r=3
r=4
$r=A_i$
$P_{\nu_i(r)}(q)$
$+\mu\sum_{r=1}^{A_i} P_{\nu_i(r)}(q)$
$-\mu n P_i(q)$
$P_i(q)$
$-\mu_d P_i(q)$
i mutates
gene duplication
time (q)
(normalization) $\mathcal{N}(q)$
$P_i(q+1)$
$\frac{W_i}{\bar{W}(q)}$
(relative fitness)
frequency change of gene pair (i,j) over time:
adjacent genes of i mutate to i
adjacent genes of i *and* j mutate to i *and* j
adjacent genes of j mutate to j
$P_{\nu_i(r)j}(q)$
$P_{\nu_i(r)\nu_j(s)}(q)$
$P_{i\nu_j(s)}(q)$
s=1
s=2
s=3
s=4
$s=A_j$
$+\mu^2\sum_{r=1}^{A_i}\sum_{s=1}^{A_j} P_{\nu_i(r)\nu_j(s)}(q)$
$+\mu\sum_{r=1}^{A_i} P_{\nu_i(r)j}(q)$
$+\mu\sum_{s=1}^{A_j} P_{i\nu_j(s)}(q)$
$P_i(q)$
$+\delta_{ij}\mu_d P_i(q)$
$P_{ij}(q)$
$-\mu 2n P_{ij}(q)$
i or j mutates
gene duplication (only if i=j)
time (q)
(normalization) $\mathcal{N}(q)$
$P_{ij}(q+1)$
$\frac{W_{ij}}{\bar{W}(q)}$
(relative fitness)

Fig. S7. Schematics of the population changes of single genes and gene pairs as described, respectively, by Eqs. **S4** (*Top*) and **S5** (*Bottom*).



Movie S1. Rotational view onto fitness landscape leading to neofunctionalization (under strong selection pressure). Gene fitness is plotted over a 2D representation of sequence space. Solid black lines indicate evolutionary trajectories of preduplication gene loci, whereas green lines indicate evolution of a gene copy arising by duplication.

Movie S1 (AVI)



Movie S2. Rotational view onto fitness landscape leading to subfunctionalization (under weak selection pressure). Gene fitness is plotted over a 2D representation of sequence space. Solid black lines indicate evolutionary trajectories of preduplication gene loci, whereas green lines indicate evolution of a gene copy arising by duplication.

Movie S2 (AVI)

Table S1. Neutral network pairs used in the present study

net 1	net 2	net 1: genes with $g = 1, 2, 3, 4, 5, 6$	net 2: genes with $g = 1, 2, 3, 4, 5, 6$	number of bridges	structure distance	prototype distance	prototype 1 (Φ_1, Φ_2)	prototype 2 (Φ_1, Φ_2)	most stable bridge ($\Phi_1 = \Phi_2$)
A	B	48, 18, 14, 22, 13, 17	20, 7, 8, 10, 7, 8	7	2	2	HPHPHPPHPHPPHPPHHH (0.998, 4.532e-05)	HPHPHHPPPHPPHPPHHH (6.438e-03, 0.955)	HPHPHHPHHHPPHPPHHH 0.444
			Fitness parameters					Fitness values in Fig. S1*C* (W_A, W_B):	
			$\tau = 1$, $\theta = 0.25$				(0.25, 4.532e-05)	(6.438e-03, 0.25)	**(0.25, 0.25)**
			$\tau = 1$, $\theta = 0.5$				(0.5, 4.532e-05)	(6.438e-03, 0.5)	**(0.444, 0.444)**
			$\tau = 1$, $\theta = 0.75$				(0.75, 4.532e-05)	(6.438e-03, 0.75)	**(0.444, 0.444)**
			$\tau = 0.3$, $\theta = 1$				(0.999, 0.024)	(0.120, 0.993)	**(0.821, 0.821)**
			$\tau = 1$, $\theta = 1$				**(0.998, 4.532e-05)**	(6.438e-03, 0.955)	(0.443, 0.443)
			$\tau = 2$, $\theta = 1$				**(0.498, 1.026e-09)**	(2.072e-05, 0.456)	(0.098, 0.098)
C	D	37, 26, 26, 25, 23, 26	23, 6, 15, 8, 11, 7	8	2	3	HPPHPPHPHPPHPHPHHH (0.998, 4.531e-05)	HPPHPPHHPPHHPHPHHH (4.287e-05, 0.944)	HPPHPPHHHPHHPHPHHH 0.438
E	F	36, 6, 28, 5, 17, 12	31, 8, 13, 10, 9, 10	5	4	4	PHPPHPHPPHPHPHHHHP (0.980, 2.998e-07)	PHPPHHHHPHPHPPHPHP (2.998e-07, 0.980)	PHPPHHHHPHPHPHHHHP 0.414
G	H	31, 8, 13, 10, 9, 10	11, 3, 9, 3, 4, 4	3	2	3	PHPPHHHHPHPHPPHPHP (0.980, 4.449e-05)	PHPPHHHHPHPHHPPHHP (3.772e-05, 0.831)	PHPPHHHHPHPHHPHHHP 0.387
I	J	29, 6, 20, 13, 16, 8	15, 27, 15, 19, 18, 11	9	3	5	HPHPHPPHPPPHHPPHHH (0.955, 2.920e-07)	HPPHHPPHHPPHPHPHHH (4.320e-05, 0.952)	HPPPHPPHHPPHPPPHHH 0.457
K	L	29, 17, 34, 11, 15, 17	22, 16, 28, 19, 16, 18	5	3	5	HHPPHPHPHPPHHPPHHH (0.946, 2.892e-07)	PHPPHPHPHPPPPHPPHH (4.323e-05, 0.952)	PHPPHPPPHPPHPHPPHH 0.287

Unless specified otherwise, simulations were started with the entire population in the prototype sequence of the first network (net 1), which is the one with a larger size in comparison with the second network (net 2). Results for network pair A and B are presented in the main text, related results for other network pairs (labeled as C and D, E and F, etc) are documented in *SI*. The numbers of genes with ground-state degeneracy $g = 1, 2, 3, 4, 5$ and 6 in the networks are given in the third and fourth columns of the table. The $g = 1$ entries are the numbers of genes with a nondegenerate ground state. The number of bridge sequences are provided in the fifth column. For every network pair tabulated here, there is at least one $g = 2$ bridge gene. Structure distance (sixth column) between the two target structures is measured by the number of different intrachain contacts between the two conformations, whereas the sequence-space distance between the prototypes (seventh column) is their Hamming distance. Also listed for each network pair are the HP model sequences for the two prototype genes and the most stable bridge gene, as well as the stabilities (fractional populations) of the structures in networks 1 and 2 (denoted respectively as Φ_1 and Φ_2) achieved by these genes (eighth, ninth, and tenth columns). The *inset* table for network pair A and B provides the fitness values (W_A, W_B) used for illustration in Fig. 1*B*. The W_A, W_B values listed here are not normalized by θ. For each τ, θ combination, the set of (W_A, W_B) that yields the highest total fitness $W = W_A + W_B$ in the model is highlighted in bold.

Table S2. Glossary of symbols used

k	genotype label for single gene i or gene pair (i, j)
i	label for a single gene, which is equivalent to a protein sequence in the model
(i, j)	label for a gene pair
W_i	fitness of gene i
W_{ij}	fitness of gene pair (i, j)
W_k	fitness of any genotype (single gene or gene pair)
$\bar{W}$	average fitness of all populated genotypes
X	set of all possible structures
X^{b}	set of beneficial target structures (only sets of two target structures are considered in this study)
X_l	any particular structure in X or X^{b}
$\Phi(X_l, i)$	stability (fractional population) of X_l relative to all possible structures of protein sequence (or gene) i
C_l	intracellular concentration of proteins folded into X_l
$W(C_l)$	fitness contribution of X_l as a function of its concentration C_l
θ	upper bound for $W(C_l)$; selection pressure
τ	deviation of $W(C_l)$ from a linear relationship with C_l in the interval $[0, \theta]$
d	dosage effect parameter; $d = 1$ for dosage increase and $d = 0$ for no dosage increase after duplication
$S_{X^{\mathrm{b}}}$	set of all single genes whose native structures belong to X^{b} (contains neutral sets for all $X_l \in X^{\mathrm{b}}$)
$D_{X^{\mathrm{b}}}$	set of all pairs of genes (i, j), with $i, j \in S_{X^{\mathrm{b}}}$
$G_{X^{\mathrm{b}}}$	union set of $S_{X^{\mathrm{b}}}$ and $D_{X^{\mathrm{b}}}$
ω	number of genes in $S_{X^{\mathrm{b}}}$
Ω	number of genotypes in $G_{X^{\mathrm{b}}}$
$P_{G_{X^{\mathrm{b}}}}$	total population (normalized to unity in this study) of all genotypes in $G_{X^{\mathrm{b}}}$
$P_i(q)$	fractional population of gene i (normalized by $P_{G_{X^{\mathrm{b}}}}$) at time q
$P_{ij}(q)$	fractional population of gene pair (i, j) (normalized by $P_{G_{X^{\mathrm{b}}}}$) at time q
q	time (discrete time steps corresponding to number of generations)
$\nu_i(r)$	label for the rth gene adjacent to gene i (in network of all genes from $S_{X^{\mathrm{b}}}$)
$\nu_j(s)$	label for the sth gene adjacent to gene j (in network of all genes from $S_{X^{\mathrm{b}}}$)
$P_{\nu_i(r)}(q)$	population of gene $\nu_i(r)$ at time q
$P_{\nu_i(r)\nu_j(s)}(q)$	population of gene pair $(\nu_i(r), \nu_j(s))$ at time q (both i and j are variables)
$P_{\nu_i(r)j}(q)$	population of gene pair $(\nu_i(r), j)$ at time q (i variable, j constant)
$P_{i\nu_j(s)}(q)$	population of gene pair $(i, \nu_j(s))$ at time q (i constant, j variable)
A_i	number of genes adjacent to gene i
A_j	number of genes adjacent to gene j
μ	rate of point mutations
μ_d	rate of gene duplications
δ_{ij}	$\delta_{ij} = 1 \Leftrightarrow i = j$ (accounts for gene duplication of i in Eq. **S5**); $\delta_{ij} = 0 \Leftrightarrow i \neq j$ (no duplication in Eq. **S5**)
n	length of gene ($n = 18$ is used for all HP model genes considered in this study)
$\mathcal{N}(q)$	normalization factor introduced at time q in Eqs. **S4** and **S5** to maintain the total population $P_{G_{X^{\mathrm{b}}}}$ at unity
$(P_{ij})_{\mathrm{st}}$	steady-state (large-time) population of genotype (i, j)
A, B	example neutral networks from the HP model
X_A,X_B	the corresponding example target structures from the HP model
π_A, π_B	the prototypes, or prototype sequences (most stable genes) in A and B, respectively
β_{AB}	the fittest (most stable) bridge sequence between neutral networks A and B
C_A, C_B	concentrations of X_A and X_B, respectively, achieved by a given genotype
W_A,W_B	fitness contributions of X_A and X_B, respectively, where $W_A = W(C_A)$ and $W_B = W(C_B)$