Supporting Information

Heo et al. 10.1073/pnas.1009392108

SI Methods

Concentration Dependence of Fitness Function: Why Cubic Root? The stoichiometric balance of protein concentrations in our model is given by the conservation equation

$$C_i = F_i + \sum_{j=1}^{7} F_{ij}. \quad \textbf{[S1]}$$

For simplicity consider a well-evolved organism where functional interactions dominate; i.e., $K_{ij}^{\mathrm{F}} \ll K_{ij}^{\mathrm{NF}}$. Then most proteins are in their functional form and we get

$$\begin{aligned} F_1 &\approx C_1 \\ F_{23} &\approx C_2 \approx C_3 \\ F_{45} + F_{64} &\approx C_4 \\ F_{45} + F_{56} &\approx C_5 \\ F_{46} + F_{56} &\approx C_6. \end{aligned} \quad \textbf{[S2]}$$

In this regime contributions to fitness function from dimers and date trimers are

$$F_{23} = \frac{1}{2}(C_2 + C_3)$$

$$F_{45}F_{56}F_{64} = \frac{1}{8}(C_4 + C_5 - C_6)(-C_4 + C_5 + C_6)(C_4 - C_5 + C_6), \quad \textbf{[S3]}$$

which explains why the cubic root in fitness function Eq. **3** of the main text is necessary to avoid bias that a priori favors one type of complex over the other.

Solution for the Law of Mass Action (LMA) Equations. For simplicity, proteins are modeled to form only monomers or dimers and all of the higher-order protein complexes are ignored in this work. The monomer concentrations of proteins, F_i were determined by solving the following seven coupled nonlinear equations of LMA (1, 2):

$$F_i = \frac{C_i}{1 + \sum_{j=1}^{N}(F_j / K_{ij})} \text{ for } i = 1, 2, \ldots, N, \quad \textbf{[S4]}$$

where N is the number of proteins in the system ($n = 7$ for the ab initio model and $n = 3{,}868$ for the proteomics simulation model) and K_{ij} defined in Eq. **8** (for the ab initio model) and Eqs. **5** and **6** (for the proteomics simulation model) of the text is the average dissociation constant of all possible interactions between proteins i and j. The concentration D_{ij} of the dimer complex between any pair of proteins is then given by the following LMA relations:

$$D_{ij} = \frac{F_i F_j}{K_{ij}}. \quad \textbf{[S5]}$$

We solved seven coupled nonlinear equations of LMA using the iteration method of refs. 1 and 2: The first iteration of F_i is calculated by substituting C_j for F_j in the right-hand side of Eq. **S1**. Each new iteration of F_i is then plugged into the right-hand side of Eq. **S1**. The iterations are repeated until the maximum relative deviation of the new values of F_i from the old ones drops below 10^{-6}.

Hydrophobicities of Evolved Proteins. To characterize the hydrophobicity of the amino acids in simulations we note that a 20 × 20 matrix of Miyazawa–Jernigan potentials, which correspond to the propensities to find interactions among 20 different types of amino acids, allow spectral decomposition with one type of eigenvalue (3, 4); i.e., an element of the matrix describing interaction energy between amino acids i and j can be presented as $E_{ij} = E_0 = \lambda q_i q_j$, where q_i is an effective hydrophobicity index of an amino acid of type i that ranges from $q_{\min} \sim 0.125$ (most hydrophilic, K) to $q_{\max} \sim 0.333$ (most hydrophobic, F). We rescaled the hydrophobicity scale to fall into a (0, 1) interval: $\tilde{q}_i = (q_i - q_{\min})/(q_{\max} - q_{\min})$. These values are presented in Table S1.

Propensities of 20 Amino Acids Constituting Functional Interfaces. We defined the propensity, Pr_a to find an amino acid type a in functional interfaces as

$$\mathrm{Pr}_a = \ln \frac{p_a}{p_a^0}, \quad \textbf{[S6]}$$

where p_a and p_a^0 are the probabilities to find an amino acid type a in sequence regions corresponding to functional interfaces and all sequence, respectively.

PPI and Protein Abundance Data for *S. cerevisiae*. We downloaded the genome-wide PPI network in baker's yeast *S. cerevisiae* from the BioGRID database (5, 6) and extracted all bait-to-prey pairs of interacting proteins detected by the affinity capture followed by mass spectrometry technique (designated as "Affinity Capture-MS" in the database). A pair of interacting proteins was then included in our "MS ≥ w" dataset if it was confirmed by at least w independent mass spectrometric experiments. We also obtained the protein expression levels of yeast proteins measured by Ghaemmaghami et al. (7). All proteins are classified with respect to their protein copy numbers using log bins. Fig. 5*D* shows the average degree of all proteins in the same concentration bin in different MS ≥ w datasets: $w = 1$ (black symbols) and 3 (red symbols).

1. Heo M, Kang L, Shakhnovich EI (2009) Emergence of species in evolutionary "simulated annealing". *Proc Natl Acad Sci USA* 106:1869–1874.
2. Maslov S, Ispolatov I (2007) Propagation of large concentration changes in reversible protein-binding networks. *Proc Natl Acad Sci USA* 104:13655–13660.
3. Miyazawa S, Jernigan RL (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.
4. Li H, Tang C, Wingreen NS (1997) Nature of driving force for protein folding: A result from analyzing the statistical potential. *Phys Rev Lett* 79:765–768.
5. Breitkreutz BJ, et al. (2008) The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res* 36(Database issue):D637–D640.
6. Stark C, et al. (2006) BioGRID: A general repository for interaction datasets. *Nucleic Acids Res* 34(Database issue):D535–D539.
7. Ghaemmaghami S, et al. (2003) Global analysis of protein expression in yeast. *Nature* 425:737–741.

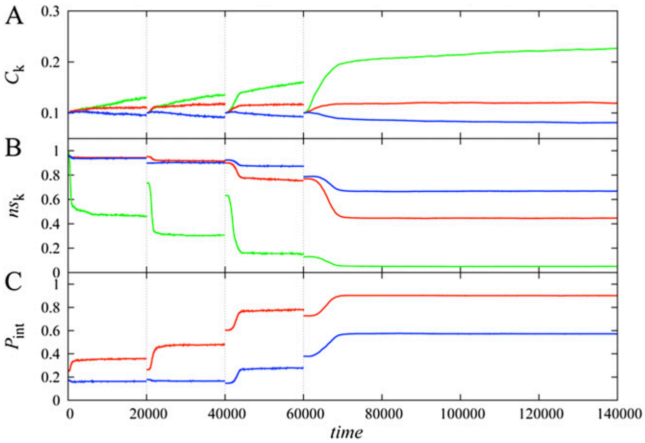

Fig. S1. Evolution of protein abundances and functional and nonfunctional protein–protein interactions. The curves represent total protein concentrations (*A*), fractional concentrations of a protein forming nonfunctional complexes (*B*), and the probability to form a functional PPI complex (*C*). The color codes represent functional monomer (protein 1, green), stable pair having one functional partner (proteins 2 and 3, red), and date triangle with two functional partners (proteins 4, 5, and 6, blue). We designed initial sequences of six cell division controlling genes (CDCG) to have highly stable structures ($P_{nat} > 0.8$) without regard for solubility of their surfaces, which resulted in mostly promiscuous nonfunctional binding of initial proteins with one another. Our population dynamics simulation consists of two parts: the first three consecutive simulations to equilibrate proteins to have proper functional interfaces depending on their functional requirements (20,000 simulation time steps each up to t = 60,000) and the last long-time production run simulation from t = 60,000 to t = 140,000, which corresponds to the simulation data presented in Fig. 2 in the main text. The vertical dotted lines partition different rounds of simulations. The seeding genome for the next round of simulation is randomly picked out of the evolved organisms in the previous round of simulation (roughly mimicking serial passage experiments), which explains the discontinuities at t = 20,000, 40,000, and 60,000. In all cases, the fraction of nonfunctional interactions of the functional monomer most drastically drops at the early stages of each round of simulation. On the other hand, the variations of nonfunctional and functional interactions of date triangle proteins are smaller than those of stable pair proteins. We averaged the curves over 100 different simulations for the first three rounds of simulations and 200 different simulations for the last round of simulation.

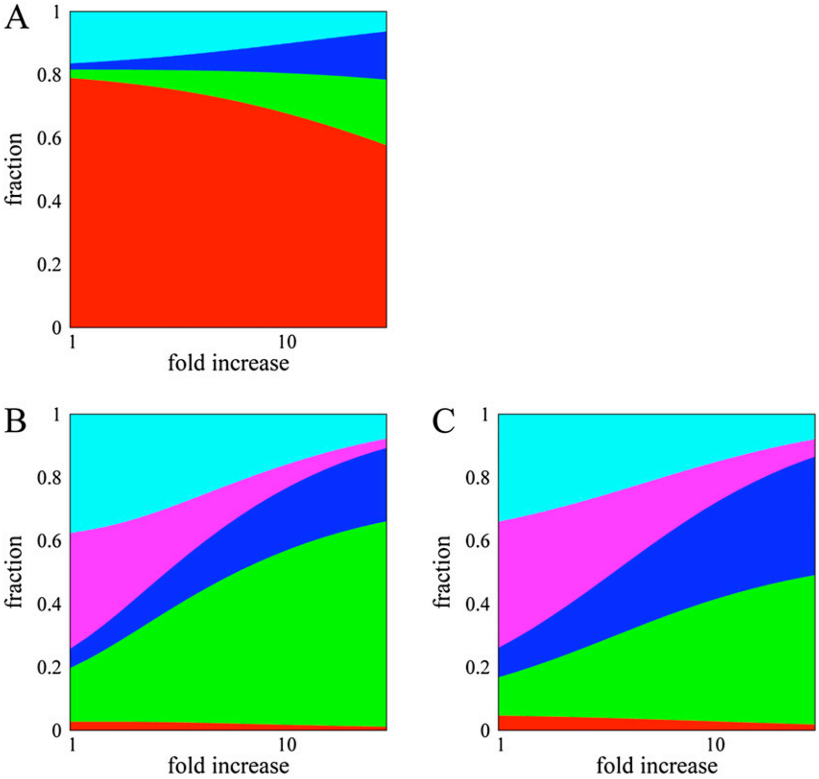

Fig. S2. Effect of dosage increase on the formation of various complexes. Colors denote various types of states of a protein: monomer (red), homodimer in head-to-head form that shares the same binding interface (green), homodimer in head-to-tail form where two participants use different binding interfaces (blue), functional heterodimer (magenta), and promiscuous complexes with a random partner (cyan). The width of each strip corresponds to the fraction of proteins in corresponding states/complexes in the cytoplasm of the model cell. The *x*-axis quantifies the level of overexpression relative to the wild-type (evolved) concentration. (*A*) Functional monomer protein. (*B*) Stable pair functional dimer proteins. (*C*) Functional dimer proteins involved in the date triangle.

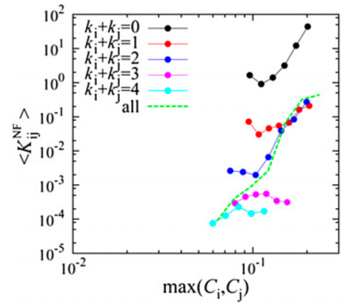

Fig. S3. Causality of the correlation between evolved PNF-PPIs and concentrations for all types of proteins. As in Fig. 4*B* we plot here dependence of K_{ij}^{NF} on concentration of interacting protein(s). However, here we present a more detailed distribution of K_{ij}^{NF} for all pairs of interacting protein types as a function of concentration of the most abundant partner, max(C_i, C_j), to distinguish between dependence on node degree and concentration. $\langle K_{ij}^{\mathrm{NF}} \rangle$ represents the average over all pairs of proteins of a particular type that fall into a given *C* bin. Different colors mark different types of pairs of interacting proteins sorted by the parameter $k_i + k_j$ – total node degree of an interacting pair of proteins *i* and *j*. For example $k_i + k_j = 0$ corresponds to homodimers of $k = 1$ proteins; $k_i + k_j = 1$ corresponds to interaction between a functional monomer and a functional dimer, $k_i + k_j = 2$ includes nonfunctional interactions (wrong surface and/or orientation) between functional dimers and interactions between functional monomers and date triangles, etc. The green line describes the average over all types as presented in Fig. 4*B*. It can be seen clearly that PNF-PPI strength is anticorrelated with protein abundances: More abundant proteins, being more ''dangerous'' to the cell in terms of their PNF-PPIs, evolve to weaken them for all interacting pairs except, perhaps, PNF-PPIs between highest node degree proteins where the ''frustration'' effect limits their ability to evolve against PNF-PPIs.

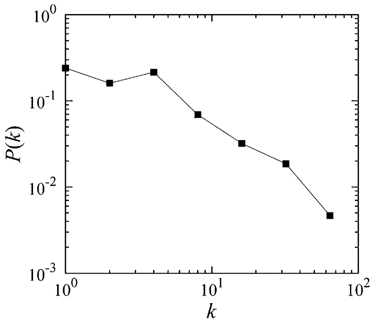

Fig. S4. The probability, *P(k)* to find a protein having node degree *k*. The artificially made true PPI network for 3,868 proteins of baker's yeast retains the scale-free property of the original one.

Table S1. Hydrophobicity of evolved proteins

	Hydrophobicity per residue		
No. of PPI partners	Functional interface	Nonbinding region	Overall sequence
$k = 0$	NA	0.29 ± 0.02	0.29 ± 0.02
$k = 1$	0.50 ± 0.02	0.29 ± 0.03	0.36 ± 0.02
$k = 2$	0.49 ± 0.03	0.30 ± 0.05	0.43 ± 0.02

Average and SDs of relative normalized hydrophobicity per residue of each sequence region are shown. The relative normalized hydrophobicity scales from 0 (most hydrophilic) to 1 (most hydrophobic). Averages and SDs are calculated over protein orthologs from 152 representative strains as described in *SI Methods*.