Supporting Information

Johnson and Hummer 10.1073/pnas.1010954108

SI Text

Z-Score Optimization. The minimum energy gap is not the only measure of fitness that could be used to optimize sequences for specificity. Z scores, for instance, were used successfully to optimize sequences against misfolding and aggregation (1). Here we found Z scores not to be sufficiently discriminating, with Z-score optimization producing sequences with much lower specificity in terms of both the average and the minimum energy gap. A high Z score, as a statistical measure, does not ensure that specific interactions are uniformly stronger than nonspecific ones. The main reason is that the average binding energy entering the Z score is dominated by the large number of very weak nonspecific interactions, in particular those associated with translated and rotated binding modes (~320 per protein pair). As a result, we found in Z-score optimizations that a few nonspecific complexes were nearly as strong as the specific ones. Our goal is to ensure that specific binding dominates over nonspecific binding, as determined from the coupled binding equilibria of the entire system. We therefore maximize the energy gap as the physical quantity that directly controls the relative amount of nonspecific binding.

Monte Carlo Sampling. In the sampling of sequence space, Monte Carlo (MC) moves consisted of point mutations and operations on pairs of specific binding partners including cyclical shifting of the pair of sequences, and partial sequence swaps that mimic V(D)J recombination (2). For the cyclical move, each residue is shifted by some number of residues h from its original location (s_i becomes s_{i+h} with periodic boundary conditions). In V(D)J recombination moves, a segment of one sequence is swapped with the corresponding segment of a binding partner. For point mutations and initialization of the random sequences, residues were selected according to the experimental residue frequency distribution. With each MC step we either mutated a single residue in 0.5–1% of the sequences (a minimum of one interfacial sequence was mutated per step), or performed a cyclical or V(D)J move on a single pair of binding partners.

To enhance global sampling in our search for sequences with a large ΔE, we used Hamiltonian replica exchange by varying the J parameter for different simulations, with the best results for values between $J = 0.1–1/k_{\mathrm{B}}T$ separated at intervals of $0.02–0.1/k_{\mathrm{B}}T$ and using 8–16 replicas. The temperature that controls the acceptance rate between different configurations is started at values around 10^{-4} and 10^{-5} and lowered 5–6 times, with each MC round consisting of 2 million mutation attempts. Additional replica exchange in the temperature variable was seen to provide only a marginal if any improvement over the original simulations.

In our interaction model we limit the size of binding interfaces to $L = 25$ residues. Increasing the interface size will increase the gap and decrease the scaling exponent, as shown for the binary model in Fig. 1*D*. However, any gains in available sequence space from larger interfaces are easily offset in a more detailed model if one considers that nonspecific binding is not restricted to the specific interfaces, but can occur anywhere at the protein surface. By trading off these two opposing effects, our model should give at least a semiquantitative description of the competition between specific and nonspecific interactions in protein–protein interaction networks.

Binding Equilibrium Calculations. To solve for the equilibrium concentrations of each protein and its complexes we use two methods, the stochastic Gillespie algorithm (3) and perturbation theory. In the Gillespie algorithm the rates of dissociation are arbitrarily set to $k_d = 1\ \mathrm{s}^{-1}$ for all complexes, and the rates of association to $k_b = k_d/K_d$ to give the proper equilibria. Initial protein concentrations (of monomers) were set at 100 nM, consistent with average protein concentrations observed in yeast (4). The volume is then chosen such that each free protein is initialized to 10,000 monomers. We run six trajectories of 10–20 s each and average the results. For each protein we then determine the ratio of the concentration of nonspecific complexes to the sum of concentrations of specific and free protein.

The specific interactions between binding partners are in general much stronger than the possible nonspecific interactions. Because of this separation in the magnitude of the dissociation constants, we can also use first order perturbation theory to accurately solve for all equilibrium concentrations. This entailed first solving the uncoupled quadratic equations for the concentrations of all specific complexes, and then correcting for the contributions of the nonspecific reactions by solving a system of linear equations. For the chain topology and the yeast networks we used only the Gillespie algorithm.

Scaling Exponent of the Energy Gap. To derive the dependence of the scaling exponent on the length of the sequences L in the binary model, we use the Hamming bound (5) for $N/2$ binary sequences, or N proteins. In the binary model the specific binding partners are always identical strings. The remaining sequence optimization against nonspecific binding is then equivalent to choosing $N/2$ points on an L-dimensional hypercube, such that the $N/2$ points are mutually as far apart as possible. An upper bound for the maximum number of proteins of length L that can be mutually separated by a Hamming distance of at least Δ is (5) $N = \frac{2^{L+1}}{\sum_{k=0}^{\lfloor(\Delta-1)/2\rfloor}\binom{L}{k}}$. The binomial sum in the denominator is equivalent to $2^L I_{1/2}(L-D,D+1)$, where $I_{1/2}$ is the incomplete beta function and $D = \lfloor\frac{\Delta-1}{2}\rfloor$. To estimate the scaling, we eliminate the floor function on Δ by using the midpoint, $D = \Delta/2 - 1$. The incomplete beta function is monotonically increasing over the range $\Delta = 0\ \mathit{to}\ L$, as is its derivative. We approximate this function by a power law $\alpha\Delta^{1/\gamma}$. To determine the dependence of the exponent γ on the length L of the sequences, we use the ratio of the function and its first derivative with respect to Δ, which results in $\gamma\Delta = \frac{I_{1/2}(L-D,D+1)}{\frac{1}{2}\frac{d}{dD}I_{1/2}(L-D,D+1)}\Big|_{D=L/4-1}$. The exponent γ is evaluated at $D = L/4 - 1$, corresponding to $\Delta = L/2$, where the scaling assumption is accurate. To simplify $I_{1/2}$ in the limit of large $L-D$ and D, we use its integral representation, $I_{1/2}(a,b) = \Gamma(a+b)\Gamma^{-1}(a)\Gamma^{-1}(b)\int_0^{1/2} t^{a-1}(1-t)^{b-1}dt$ and approximate the integrand $t^{L-D-1}(t-1)^D$ with an exponential $2^{1-L}\exp[2(L-2D-1)(t-1/2)]$ that is readily integrated. For the gamma functions, we use Stirling's approximation. For large L, the exponent in the power law becomes $\gamma = (\frac{\ln 3}{4}L + \frac{4+\ln 3}{3})^{-1}$, scaling as $1/L$.

1. Deeds EJ, Ashenberg O, Gerardin J, Shakhnovich EI (2007) Robust protein-protein interactions in crowded cellular environments. *Proc Natl Acad Sci USA* 104:14952–14957.
2. Alt FW, et al. (1992) VDJ Recombination. *Immunol Today* 13(8):306–314.
3. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81:2340–2361.

4. Ghaemmaghami S, et al. (2003) Global analysis of protein expression in yeast. *Nature* 425:737–741.

5. MacWilliams FJ, Sloane NJA (1977) *The Theory of Error-Correcting Codes* (Elsevier, Amsterdam).

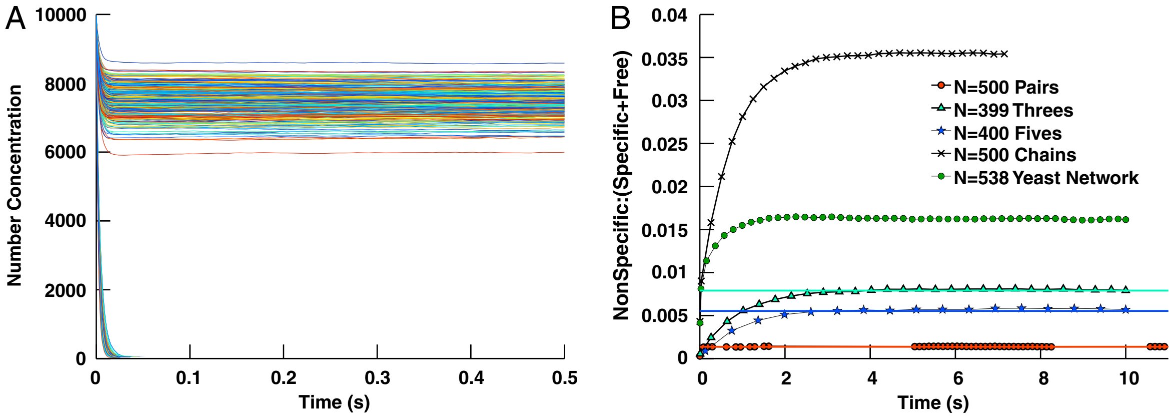

Fig. S1. Convergence of the Gillespie algorithm simulations. (*A*) Number concentration of free proteins as a function of time for a system of $N = 400$ proteins in the Fives network topology, simulated with the Gillespie algorithm. Each colored line is a distinct protein. The hub proteins are rapidly consumed in complexes with their partners, and hence their free protein concentrations approach zero. The four partners of these hub proteins must compete for the hub protein, and therefore about 25% of each type of protein forms complexes, and most of the remainder, 75%, remain free in solution. (*B*) Concentration of nonspecific complexes, normalized by the specific complex plus free protein concentrations, as a function of time simulated with the Gillespie algorithm. Results are shown for systems with different topologies and $N \approx 500$. Convergence is achieved for all systems after about 3–4 s. The data in *A* correspond to the ratio shown with blue stars. For the Pairs, Threes and Fives topology we also plot with horizontal lines the value calculated using perturbation theory, which is in excellent agreement for all three topologies.

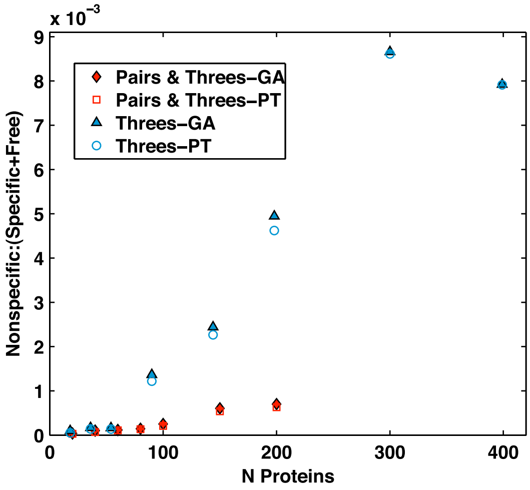

Fig. S2. Comparison of the average concentration of nonspecific complexes normalized by the sum of the concentrations of the specific complex and free protein, calculated with the Gillespie algorithm (GA, filled symbols) and perturbation theory (PT, open symbols). We plot results for systems optimized with the Pairs and Threes topology as well as the Threes topology.

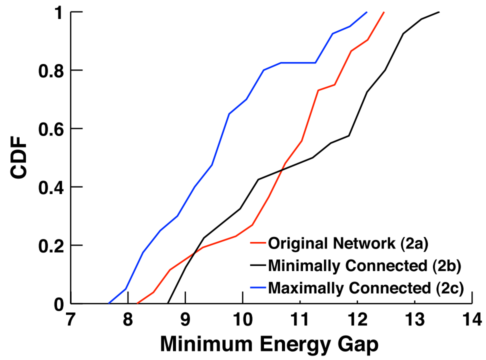

Fig. S3. Cumulative distribution function (CDF) of the probability of observing a minimum energy gap for each of the interfaces in the yeast interaction network fragment. The red line is the CDF for the original network connectivity of 52 interfaces (Fig. 2*A*). The black and the blue lines are for the modified 40 interface networks with minimal number of chains (Fig. 2*B*) and maximal number of chains (Fig. 2*C*), respectively. In all three networks, the highly connected interfaces are responsible for the smallest ΔE values. The network with the largest ΔE (i.e., highest specificity) is the minimally connected network, whereas the maximally connected network has the smallest ΔE (lowest specificity).

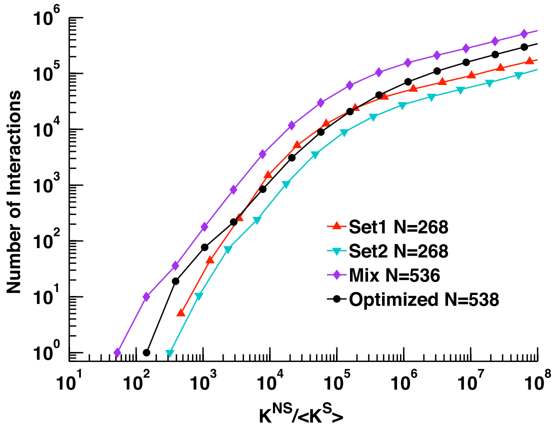

Fig. S4. Effect on nonspecific binding from mixing of subcellular compartments. The curves show the distributions of the dissociation constants K^{NS} of nonspecific complexes in the replicated yeast network. K^{NS} is normalized by the mean specific dissociation constant, $\langle K^S \rangle$, obtained from the average specific binding energy. Results are shown for the two best, independently optimized sequence sets with $N = 268$ interfaces each (red and blue triangles), and for the combined set of $2N$ interfaces without further optimization (purple diamonds). Additional sequence optimization (black circles) shifts the curve to the right, thus reducing the number of competitive nonspecific complexes.

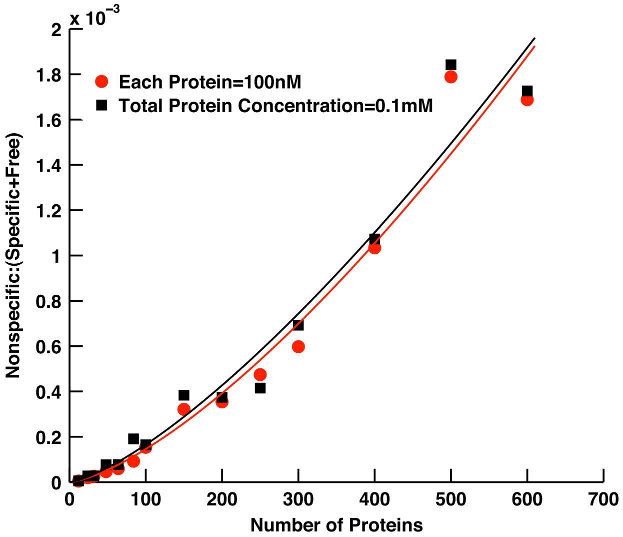

Fig. S5. Comparison of nonspecific complex formation with different protein concentrations. For the Pairs network we calculated the fraction of nonspecific complexes formed at equilibrium for the case that each protein has a total concentration of 100 nM irrespective of the number N of proteins (red circles), and the case that the total protein concentration is 0.1 mM, divided equally among the N proteins (black squares). For the second case, the total concentration of each protein is thus $(0.1\ \text{mM})/N$. The red and black lines are guides to the eye that represent the best fit to the data using the function aN^b, where a and b are fit parameters.