Simulated Evolution of Protein-Protein Interaction Networks with Realistic Topology

G. Jack Peterson[1]*, Steve Pressé[2], Kristin S. Peterson[3], Ken A. Dill[4]

[1] Biophysics Graduate Group, University of California San Francisco, San Francisco, California, United States of America, [2] Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California, United States of America, [3] Department of Forest Ecosystems and Society, Oregon State University, Corvallis, Oregon, United States of America, [4] Laufer Center for Physical and Quantitative Biology, Stony Brook University, New York, United States of America

Abstract

We model the evolution of eukaryotic protein-protein interaction (PPI) networks. In our model, PPI networks evolve by two known biological mechanisms: (1) Gene duplication, which is followed by rapid diversification of duplicate interactions. (2) Neofunctionalization, in which a mutation leads to a new interaction with some other protein. Since many interactions are due to simple surface compatibility, we hypothesize there is an increased likelihood of interacting with other proteins in the target protein's neighborhood. We find good agreement of the model on 10 different network properties compared to high-confidence experimental PPI networks in yeast, fruit flies, and humans. Key findings are: (1) PPI networks evolve modular structures, with no need to invoke particular selection pressures. (2) Proteins in cells have on average about 6 degrees of separation, similar to some social networks, such as human-communication and actor networks. (3) Unlike social networks, which have a shrinking diameter (degree of maximum separation) over time, PPI networks are predicted to grow in diameter. (4) The model indicates that evolutionarily old proteins should have higher connectivities and be more centrally embedded in their networks. This suggests a way in which present-day proteomics data could provide insights into biological evolution.

Citation: Peterson GJ, Pressé S, Peterson KS, Dill KA (2012) Simulated Evolution of Protein-Protein Interaction Networks with Realistic Topology. PLoS ONE 7(6): e39052. doi:10.1371/journal.pone.0039052

Editor: Stefano Boccaletti, Technical University of Madrid, Italy

Received April 16, 2012; **Accepted** May 15, 2012; **Published** June 29, 2012

Funding: GJP was funded by a National Defense Science and Engineering Graduate Fellowship from the Department of Defense, SP was funded by a Sandler Blue Sky Award, KSP was funded by a Long Term Ecological Research grant from the National Science Foundation, and KAD was funded by the National Science Foundation and National Institutes of Health GM 34993. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jack@tinybike.net

Introduction

We are interested in the evolution of protein-protein interaction (PPI) networks. PPI network evolution accompanies cellular evolution, and may be important for processes such as the emergence of antibiotic resistance in bacteria [1,2], the growth of cancer cells [3], and biological speciation [4–6]. In recent years, increasingly large volumes of experimental PPI data have become available [7–10], and a variety of computational techniques have been created to process and analyze these data [11–18]. Although these techniques are diverse, and the experimental data are noisy [19], a general picture emerging from these studies is that the evolutionary pressures shaping protein networks are deeply interlinked with the networks' topology [20]. Our aim here is to construct a minimal model of PPI network evolution which accurately captures a broad panel of topological properties.

In this work, we describe an evolutionary model for eukaryotic PPI networks. In our model, protein networks evolve by two known biological mechanisms: (1) a gene can duplicate, putting one copy under new selective pressures that allow it to establish new relationships to other proteins in the cell, and (2) a protein undergoes a mutation that causes it to develop new binding or new functional relationships with existing proteins. In addition, we allow for the possibility that once a mutated protein develops a new relationship with another protein (called the target), the mutant protein can also more readily establish relationships with other proteins in the target's neighborhood. One goal is to see if random changes based on these mechanisms could generate networks with the properties of present-day PPI networks. Another goal is then to draw inferences about the evolutionary histories of PPI networks.

Results

We represent a PPI network as a graph. Each node on the graph represents one protein. A link (edge) between two nodes represents a physical interaction between the two corresponding proteins. The links are undirected and unweighted. To model the evolution of the PPI graph, we simulate a series of steps in time. At time t, one protein in the network is subjected to either a gene duplication or a neofunctionalizing mutation, leading to an altered network by time $t+\Delta t$. We refer to this model as the DUNE (DUplication & NEofunctionalization) model.

Gene Duplication

One mechanism by which PPI networks change is gene duplication (DU) [21–23]. In DU, an existing gene is copied, creating a new, identical gene. In our model, duplications occur at a rate d, which is assumed to be constant for each organism. All genes are accessible to duplication, with equal likelihood. For simplicity, we assume that one gene codes for one protein. One of the copies continues to perform the same biological function and remains under the same selective pressures as before. The other copy is superfluous, since it is no longer essential for the functioning of the cell [24].

The superfluous copy of a protein/gene is under less selective pressure; it is free to lose its previous function and to develop some other function within the cell. Due to this reduced selective pressure, further mutations to the superfluous protein are more readily accepted, including those that would otherwise have been harmful to the organism [25,26]. Hence, a superfluous protein diverges rapidly after its DU event [27,28]. This well-known process is referred to as the *post-duplication divergence*. Following [29], we assume that the link of each such superfluous protein/gene to its former neighbors is deleted with probability ϕ. The post-duplication divergence tends to be fast; for simplicity, we assume the divergence occurs within the same time step as the DU. The divergence is asymmetric [30,31]: one of the proteins diversifies rapidly, while the other protein retains its prior activity. We delete links from the original or the duplicate with equal probability because the proteins are identical. As discussed in the supporting information (SI), this is closely related to the idea of *subfunctionalization*, where divergence freely occurs until redundancy is eliminated (see SI text in File S1). In our model, ϕ is an adjustable parameter.

In many cases, the post-duplication divergence results in a protein which has lost all its links. These 'orphan' proteins correspond to silenced or deleted genes in our model. As discussed below, our model predicts that the gene loss rate should be slightly higher than the duplication rate in yeast, and slightly lower in flies and humans.

We simulate a gene duplication event at time t as follows:

1a. Duplicate a randomly-chosen gene with probability $d\Delta t$.

2a. Choose either the original (50%) or duplicate (50%), and delete each of its links with probability ϕ.

3a. Move on to the next time interval, time $t+\Delta t$.

Neofunctionalization

Our model also takes into account that DNA can be changed by random mutations. Most such mutations do not lead to changes in the PPI network structure. However, some protein mutations lead to new interactions with some other protein (which we call the *target protein*). The formation of a novel interaction is called a *neofunctionalization* (NE) event. NE refers to the creation of new interactions, not to the disappearance of old ones. Functional deletions tend to be deleterious to organisms [32]. We do not account for loss-of-function mutations (link deletions) except during post-duplication divergence because damaged alleles will, in general, be eliminated by purifying selection. In our model, NE mutations occur at a rate μ, which is assumed to be constant. All proteins are equally likely to be mutated.

How does the mutated protein choose a target protein to which it links? We define a probability q that any protein in the network is selected for receiving the new link from the mutant protein. To account for the possibility of homodimerization, the mutated protein may also link to itself [24,33]. Random choice dictates that $q=1/N$ (see SI).

Many PPI's are driven by a simple geometric compatibility between the surfaces of the proteins [34]. The simplest example is the case of PPI's between flat, hydrophobic surfaces [35], a type of interaction which is very common [36]. These PPI's have a simple planar interface, and the binding sites on the individual proteins are geometrically quite similar to one another. One consequence of these similar-surface interactions is that if protein A can bind to proteins B and C, then there is a greater-than-random chance that B and C will interact with each other. We refer to this property as *transitivity*: if A binds B, and A binds C, then B binds C. The number of triangles in the PPI network should correlate roughly with transitivity. As discussed below, the number of triangles (as quantified by the global clustering coefficient) is about 45 times higher in real PPI networks than in an equally-dense random graph. This suggests that transitivity is quite common in PPI networks. Another source of transitivity is gene duplication. If A binds B, then A is copied to create a duplicate protein A', then A' will (initially) also bind B. If A interacts with A', then a triangle exists. However, duplication is unlikely to be the primary source of transitivity; recent evidence shows that, due to the post-duplication divergence, duplicates tend to participate in fewer triangles than other proteins [37].

A concrete example of transitivity is provided by the evolution of the retinoic acid receptor (RAR), an example of neofunctionalization which has been characterized in detail [38]. Three paralogs of RAR exist in vertebrates (RARα, β, and γ), as a result of an ancient duplication. The interaction profiles of these proteins are quite different. Previous work indicates that RARβ retained the role of the ancestral RAR [38], while RARα and γ evolved new functionality. RARα has several interactions not found in RARβ. RARα has novel interactions with a histone deacetylase (HDAC3) as well as seven of HDAC3's nearest-neighbors (HDAC4, MBD1, Q15959, NRIP1, Q59FP9, NR2E3, GATA2). None of these interactions are found in RARβ. The probability that all of these novel interactions were created independently is very low. RARα has 65 known PPI's and HDAC3 has 83, and the present-day size of the human PPI network is a little over 3000 proteins. Therefore, the chance of RARα randomly evolving novel interactions with 7 of HDAC3's neighbors is less than 1 in a billion. This strongly suggests that when a protein evolves an interaction to a target, it has a greater-than-random chance of also linking to other, neighboring proteins.

How do similar-surface interactions affect the evolution of PPI networks? First, consider how an interaction triangle would form. Suppose proteins A and B bind due to physically similar binding sites. Protein X mutates and evolves the capacity to bind A. There is a reasonable chance that X has a surface which is similar to both A and B. If so, protein X is likely to also bind to B, forming a triangle. Denote the probability that two proteins interact due to a simple binding site similarity by a. The probability that A binds B (and X binds A) in this manner is a. Assuming these probabilities are identical and independent, the probability that X binds B is a^2.

So far, we have discussed transitivity as it affects the PPI's in which protein A is directly involved (A's first-neighbors). We now introduce a third protein to the above example, resulting in a chain of interactions: protein A binds B, B binds C, but C does not bind A. Protein X mutates and gains an interaction with A (with probability a^2). What is the probability that X will also bind C? The probability that B binds C due to surface similarity is a. Thus, X will bind C (A's second-neighbor) with probability a^3. In general, the probability that X will bind one of A's j^{th} neighbors is a^{j+1}. We refer to this process as *assimilation*, and the 'assimilation parameter' a is a constant which varies between species. As discussed in SI, it is primarily mutliple-partner proteins which bind to their partners at different times and/or locations which are affected by this process; consequently, at most one link is created by assimilation at the first-neighbor level, second-neighbor level, etc. Assimilation is assumed to act on a much shorter time scale than duplication and neofunctionalization; in our model, it is instantaneous.

Our hypothesized assimilation mechanism makes several predictions that could be tested experimentally: (1) the probability of a protein assimilating into a new pathway should be a^2 (at the first-neighbor level), a^3 (at the second-neighbor level), and so on, where a is a constant which varies between species; (2) weak, nonspecific binding and planar interfaces should be overrepresented in interaction triangles (and longer cycles) between non-duplicate proteins; (3) competitive inhibitors should be overrepresented in interaction triangles; and (4) domain shuffling should be associated with assimilation. (See SI for discussion of (3) and (4).).

We simulate a neofunctionalization event at time t as follows:

1b. Mutate a randomly-chosen gene with probability $\mu\Delta t$.
2b. Link to a randomly-chosen target protein.
3b. Add a second link to one of the target's first-neighbor proteins, chosen randomly, with probability a^2.
4b. Add a link to one of the target's second-neighbor proteins, with probability a^3, etc.
5b. Move on to the next time interval, time $t+\Delta t$.

Model Simulation and Parameters

A flowchart of how PPI networks evolve in our model is shown in Figure 1. To simulate the network's evolution, one of the two mechanisms above is used at each time step, using [39]. We call each possible time series a *trajectory*. We begin each trajectory starting from two proteins sharing a link (the simplest configuration that is still technically a network). Each simulated trajectory ends when the model network has grown to have the same total number of links, K, as found in the experimental data, K_{data}. Here, we perform sets of simulations for three different organisms: yeast (*Saccharomyces cerevisiae*), fruit flies (*Drosophila melanogaster*), and humans (*Homo sapiens*). Because evolution is stochastic, there are different possible trajectories, even for identical starting conditions and parameters. We simulated 50 trajectories for each organism. Our figures show the median values of each feature as a heavy line, and individual trajectories as light lines.

For a given data set, the number of links (K_{data}) is known. We estimate the duplication rate d from literature values. There have been several empirical estimates of duplication rates, mostly falling within an order of magnitude of each other [27,40–42,42–45]. We averaged together the literature values to estimate d for each species (Table 1).

The quantity μ is not as well known. Its value relative to d has been the topic of considerable debate [24,46–48]. Although, in principle, μ is a measurable quantity, it has proven difficult to obtain an accurate value, in part because the fixation rate of neofunctionalized alleles varies with population size [49,50]. In the absence of a consensus order-of-magnitude estimate, in our model, we treat μ as a fitting parameter. Consistent with the findings of [51] and [46], our best-fit values of μ are within an order of magnitude of each other for yeast, fruit fly, and human networks. Best-fit parameter values are given in Table 1.

Present-day Network Topology

One test of an evolutionary model is its predictions for present-day PPI network topologies. Current large-scale PPI data sets have a high level of noise, resulting in significant problems with false positives and negatives [19,52]. To mitigate this, we compare only to 'high-confidence' experimental PPI network data gathered in small-scale experiments (see Methods). We computed 10 topological features, quantifying various static and dynamic aspects of the networks' global and local structures: degree, closeness, eigenvalues, betweenness, modularity, diameter, error tolerance, largest component size, clustering coefficients, and assortativity. 8 of these properties are described below (see SI for others).

The *degree* k of a node is the number of links connected to it. For protein networks, a protein's degree is the number of proteins with which it has direct interactions. Some proteins interact with few other proteins, while other proteins (called 'hubs') interact with many other proteins. Previous work indicates that hubs have structural and functional characteristics that distinguish them from non-hubs, such as increased proportion of disordered surface residues and repetitive domain structures [53]. The high degree of a protein hub could indicate that protein has unusual biological significance [54]. The network's overall link density is described by its mean degree, $\langle k \rangle$ (Table 2). The *degree distribution* $p(k)$ is the probability that a protein will have k links. PPI networks have a few hub proteins and many relatively isolated proteins. The heavy tail of the degree distribution shows that PPI networks have significantly more hubs than random networks have. Simulated and experimental degree distributions are compared in Figure 2. (For quantitative comparisons, see SI.).

Component refers to a set of reachable proteins. If any protein is reachable from any other protein (by hopping from neighbor to neighbor), then the network only has one component. If there is no path leading from protein A to B, then A and B are in different components. The fraction of nodes in the largest component (f_1) is a measure of network fragmentation (Table 2 and Figure S3). Note that, although silent genes (proteins with no links) exist in real systems, these genes do not appear in data sets consisting only of PPI's. Therefore, calculations of f_1 for all models exclude orphan proteins (proteins with $k=0$).

Gene loss, the silencing or deletion of genes, is known to play an important role in evolution. The loss of a functioning gene will damage an organism, making the gene loss unlikely to be passed on. The exception is if the gene is redundant. Consistent with this reasoning, evidence suggests that many gene loss events are losses of one copy of a duplicated gene [30,55]. Although empirical estimates of the gene loss rate varied considerably, a consistent finding across several studies is that the rates of gene duplication and loss are of the same order-of-magnitude [27,41,44]. This broad picture is in good agreement with our model. In our model, a gene is considered lost when it has degree zero. Our model predicts that the ratio of orphan to non-orphan proteins is 1.6 ± 0.4 in yeast, 0.58 ± 0.06 in flies, and 0.67 ± 0.09 in humans. The gene loss rate has been previously estimated to be about half the duplication rate in both flies and humans [27,44], consistent with our model's prediction.

The *distance* between nodes i and j is defined as the number of node-to-node steps that it takes along the shortest path to get from node i to j. The *closeness centrality* of a node i, ℓ_i, is the inverse of the average distance from node i to all other nodes in the same component. The *diameter*, D, of a network is the longest distance in the network. Simulated closeness distributions are compared to experiments in Figure 3. Interestingly, proteins have about 'six degrees of separation', similar to social networks [56,57]. The closeness distributions $p(\ell)$ have peaks around $1/\ell \approx 5-7$.

Another property of a network is its *modularity* [58]. Networks are modular if they have high densities of links (defining regions called modules), connected by lower densities of links (between modules). One way to quantify the extent of modular organization in a network is to compute the modularity index, Q [59,60]:

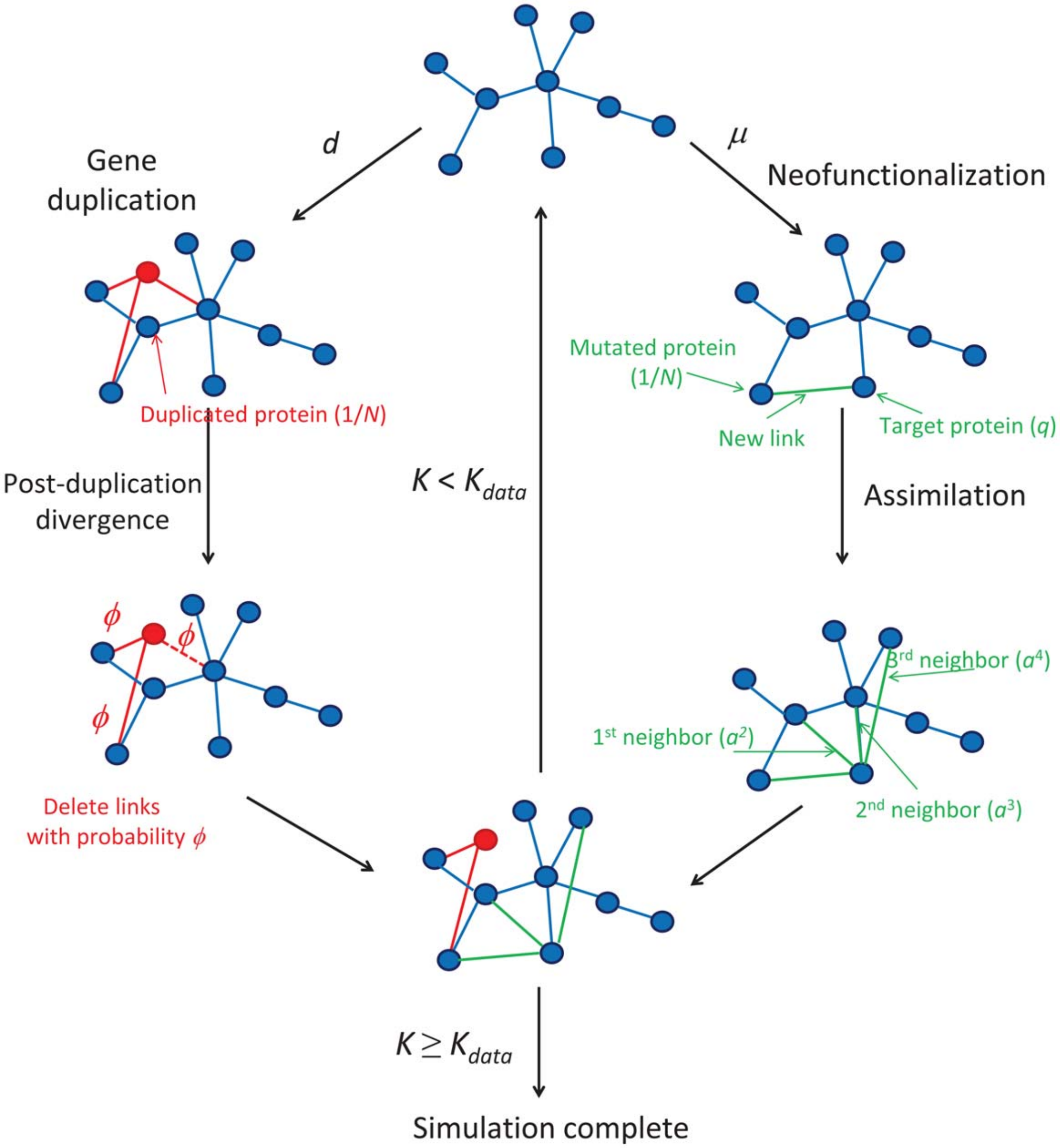

Figure 1. DUNE model flowchart. At each time step, the simulated network undergoes a duplication or neofunctionalization event. Red nodes/links indicate nodes/links that have been created by duplication during the current time step. Green links indicate links that have been created by neofunctionalization during the current time step. A dashed line indicates a duplicated link that has been deleted during the post-duplication divergence. Only 3 neighbors are shown for the assimilation mechanism; however, the actual simulations included up to 20th neighbors. The simulated network evolves until its number of links (K) meets or exceeds the number of links in the data (K_{data}).
doi:10.1371/journal.pone.0039052.g001

$$Q \equiv \frac{1}{K}\sum_{i,j}^{N}\left(A_{ij}-\frac{k_i k_j}{K}\right)\delta(u_i,u_j), \qquad (1)$$

where k_i and k_j are the degrees of nodes i and j, u_i and u_j denote the modules to which nodes i and j belong, $\delta(u_i,u_j)=1$ if $u_i=u_j$ and $\delta(u_i,u_j)=0$ otherwise, and $A_{ij}=1$ if nodes i and j share a link, and $A_{ij}=0$ otherwise. Q quantifies the difference between the actual within-module link density to the expected link density in a randomly connected network. Q ranges between -1 and 1; positive values of Q indicate that the number of links within modules is greater than random. The numerical value of Q required for a network to be considered 'modular' depends on the number of nodes and links and method of computation. To calibrate baseline Q values given our particular network data, we used the null model described in [61]. Our non-modular baseline values are $Q=0.603$ for the human PPI net, $Q=0.590$ for yeast, and $Q=0.722$ for flies (see SI). As shown in Table 2, PPI networks are highly modular, and our simulated Q values are in good agreement with those of experimental data.

The *clustering coefficient*, C_i, for a protein i, is a measure of mutual connectivity of the neighbors of protein i. C_i is defined as the ratio of the actual number of links between neighbors of protein i to the maximum possible number of links between them,

Table 1. Network sizes and model parameters.

	N_{data}	K_{data}	d	μ	ϕ	a
Yeast	2170	3819	0.01	7.86×10^{-4}	0.555	0.690
Fly	878	1140	0.0014	5.89×10^{-4}	0.866	0.546
Human	3165	5547	0.0037	7.62×10^{-4}	0.652	0.727

N and K are the numbers of proteins and links, respectively. (K_{data} is used to stop the simulation. N_{data} is not used as a constraint.) d and μ have units of per gene per million years (Myr). ϕ and a are probabilities (unitless). K_{data} and d are constraints from the data, while μ, ϕ, and a are adjustable parameters. We used Monte Carlo simulations to optimize the parameter values, by minimizing the total symmetric mean absolute percentage error values of the simulated versus the experimental data (see SI). Our values of μ are substantially lower than d because μ is the rate of mutations leading to the creation of a new PPI (rather than being a simple mutation rate, which would be much higher).
doi:10.1371/journal.pone.0039052.t001

$$C_i = \frac{\#\ \text{edges between neighbors of node}\ i}{k_i(k_i-1)}. \qquad (2)$$

In a PPI network, clustering is thought to reflect the high likelihood that proteins of similar function are mutually connected [62]. The average (or global) clustering coefficient, $\langle C\rangle$, quantifies the extent of clustering in the network as a whole. As shown in Table 2, PPI networks have large global clustering coefficient values; the yeast PPI network, for example, has a value of $\langle C\rangle$ which is 45 times higher than that of a random graph of equivalent link density. In flies and humans, our simulated networks have $\langle C\rangle$ values in excellent agreement with the data; in yeast, our predicted value is slightly low.

A network is said to be 'hierarchically clustered' if the clustering coefficient and degree obey a power-law relation, $C\tilde{}k^{-\xi}$ [63] (Figure S1), indicating that nodes are organized into small-scale modules, and the small-scale modules are in turn organized into larger-scale modules following the same pattern [64]. By plotting each node's clustering coefficient against its degree, we observed a trend consistent with hierarchical clustering, although data in the tail is very limited.

The *betweenness* of a node measures the extent to which it 'bridges' between different modules. *Betweenness centrality*, b, is defined as:

$$b_i \equiv \frac{\#\ \text{shortest paths passing through node}\ i}{\#\ \text{total shortest paths}}. \qquad (3)$$

Betweenness has been proposed as a uniquely functionally-relevant metric for PPI networks because it relates local and global topology. It has been argued that knocking out a protein that has high betweenness may be more lethal to an organism than knocking out a protein of high degree [65]. Betweenness distributions are shown in Figure 4.

If a network's well-connected nodes are mostly attached to poorly-connected nodes, the network is called *disassortative*. A simple way to quantify disassortativity is by determining the median degree of a protein's neighbors (n) as a function of its degree (k). Previous work has found that yeast networks are disassortative [61]. It has been argued that disassortativity is an essential feature of PPI network evolution, and recent modeling efforts have heavily emphasized this feature [66,67]. However, it was noted by [68] that disassortativity may simply be an artifact of the yeast two-hybrid technique, and [69] pointed out that this trend is quite different among different yeast datasets, and in some cases is completely reversed, resulting in *assortative* mixing, where high degree proteins prefer to link to other high-degree proteins. As shown in Figure 5 and Table S1, the empirical data shows no evidence of disassortativity in flies or humans, and even the trend in yeast is quite weak. This conclusion is based solely on analysis of the empirical data, and casts further doubt on the role of disassortative mixing in PPI network evolution.

Table 2. Comparison of network features.

	Q	D	f_1	$\langle C\rangle$	$\langle k\rangle$
Yeast data	**0.75**	**15**	**0.89**	**0.09**	**3.65**
DUNE	0.74(7)	17(6)	0.8(1)	0.041(9)	4.0(8)
Vázquez	0.80(4)	21(5)	0.2(1)	0.045(5)	2.6(4)
Berg	0.518(4)	12.0(7)	0.990(3)	0.0027(9)	4.10(3)
RG	0.910(3)	36(3)	0.987(6)	0.475(8)	5.31(8)
MpK	0.58(6)	24(5)	1.000(2)	0.08(3)	4.4(6)
ER	0.588(8)	13.0(9)	0.995(2)	0.002(1)	3.5(6)
Fly data	**0.86**	**23**	**0.73**	**0.10**	**2.93**
DUNE	0.82(2)	20(2)	0.81(3)	0.09(1)	2.36(9)
Human data	**0.75**	**15**	**0.88**	**0.08**	**3.69**
DUNE	0.74(6)	17(2)	0.88(4)	0.09(1)	3.7(4)

Modularity Q, diameter D, fraction of nodes in the largest component f_1, global clustering coefficient $\langle C\rangle$, and $\langle k\rangle$ is the average degree of proteins the largest component. 'Data' is the empirical data, 'DUNE' is the model described here, 'Vázquez' is the duplication-only model of [29], 'Berg' is the link dynamics model [85], 'RG' is random geometric [89], 'MpK' is the physical desolvation model presented in [52], and 'ER' is an Erdös-Rényi random graph [90]. Simulated values are the median ($\pm$ standard deviation) over 50 simulations. (See SI for details of each model's setup and optimization.).
doi:10.1371/journal.pone.0039052.t002

Comparisons of simulated and experimental eigenvalue spectra and error tolerance curves are shown in SI (Figures S7 and S8). As discussed in SI, the various per-node network properties we have analyzed are largely uncorrelated (Figure S9).

Evolutionary Trajectories

We now consider the question of how PPI networks evolve in time. The present-day networks show a rich-get-richer structure: PPI networks tend to have both more well-connected nodes and more poorly connected nodes than random networks have. In our model, the rich-get-richer property has two bases: duplication and assimilation. The equal duplication chance per protein means the probability for a protein with k links to acquire a new link via duplication of one of its interaction partners is proportional to k. Likewise, the probability of a protein to receive a link from the first-neighbor assimilation probability a is proportional to its degree k. 'Rich' proteins get richer because the probability of acquiring new links rises with the number of existing links.

First, we discuss two dynamical quantities for which experimental evidence exists: the rate of gene loss, and the relation between a protein's age and its centrality. Gene losses in our model correspond to 'orphan' proteins which have no interactions with other proteins. As shown in Figure S3, the fraction of orphan proteins grows quickly at first, then levels off. This is consistent with the findings of [44]: in humans, while the overall duplication rate is higher than the loss rate, when only data from the past 200 Myr are considered, the loss rate is slightly higher than the

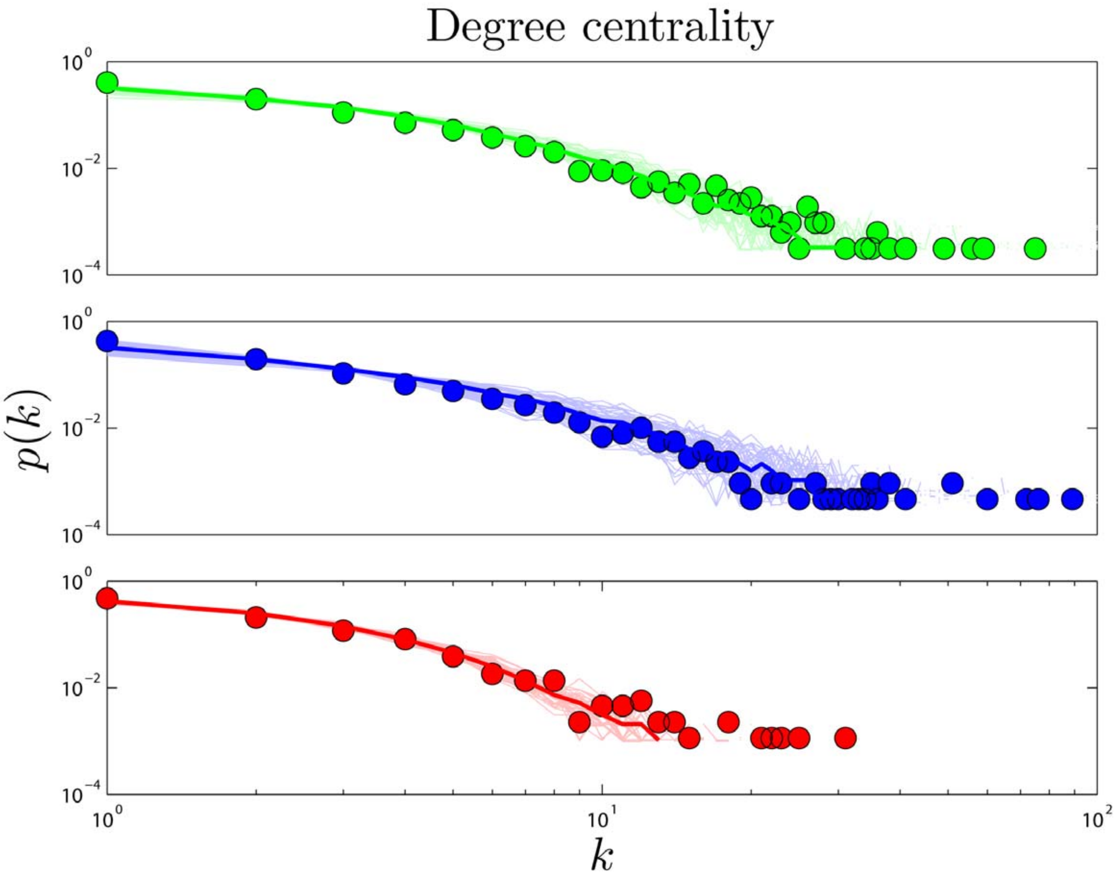

Figure 2. Degree centrality. Degree (k) distributions in human (green), yeast (blue), and fly (red). Heavy lines are the median values from 50 simulations, and light lines are results of individual simulations. Points represent high-confidence empirical data for each organism (see Methods). Unless otherwise noted, color coding in the same in all plots. Quantitative comparisons between simulation and experiment (for DUNE and several other models) are detailed in SI.
doi:10.1371/journal.pone.0039052.g002

duplication rate. In our model, after the initial rapid expansion, the rate of gene loss stabilizes relative to the duplication rate.

We define the 'age' of a protein in our simulation according to the order in which proteins were added to the network. Our model shows that a protein's age correlates with certain network properties. Consistent with earlier work [70–73], we find that older proteins tend to be more highly connected. We plotted the 'age index' of a protein (the time step at which the protein was introduced) versus its centrality scores. As shown in Figure S2, the age index negatively correlates with degree, betweenness, and closeness centralities: older proteins tend to be more central than younger proteins. Figure S2 shows our model's prediction that a protein's age correlates with degree, betweenness, and closeness centrality. We confirmed this prediction by following the evolutionary trajectories of individual proteins (Figure S4). These results are consistent with the eigenvalue-based aging method described in [73] (Figure S5). Phylogenetic protein age estimates indicate that older proteins tend to have a higher degree [70,73], which our model correctly predicts. Interestingly, the eigenvalue-based scores are only modestly correlated with other centrality scores (0.36 degree, 0.47 betweenness, and 0.10 closeness correlations). Using the eigenvalue method in tandem with our centrality-based method could provide stronger age-discriminating power for PPI networks than either method alone.

The correlation between centrality and age suggests that static properties of present-day networks may be used to estimate relative protein ages. Suppose each normalized centrality score ($k' \equiv k/\max(k)$, $\ell' \equiv \ell/\max(\ell)$, $b' \equiv b/\max(b)$) represents a coordinate in a 3-D 'centrality space'. We can then define a composite centrality score (S) as $S^2 \equiv (k')^2 + (\ell')^2 + (b')^2$.

Do older proteins typically have different functions than newer proteins? We classified *S. cerevisiae* proteins using the GO-slim gene ontology system in the Saccharomyces Genome Database. As shown in Figure S6, GO-slim enrichment profiles were somewhat different between the oldest and youngest proteins (as measured by their S values). Several categories which were more enriched for the oldest proteins were the cell cycle, stress response, cytoskeletal and cell membrane organization, whereas younger proteins were overrepresented in several metabolic processes. Overall, the differences were not dramatic, suggesting that cellular processes generally require both central and non-central proteins to function. Consistent with this, ancient proteins tend to be centrally located with modules, as their betweenness values gradually decline over time (Figure S4). The roughly linear relation between degree and betweenness also suggests that ancient proteins do not occupy structurally 'special' positions within the network, such as stitching together separate modules (Table S1 and Figure S10). This may indicate that modules tend to accumulate around the most ancient proteins, which act as a sort of nucleus. Thus, ancient proteins are involved in all kinds of pathways, because they have each nucleated their own pathway.

In contrast to the two dynamical quantities discussed so far, most structural properties of PPI networks have only been

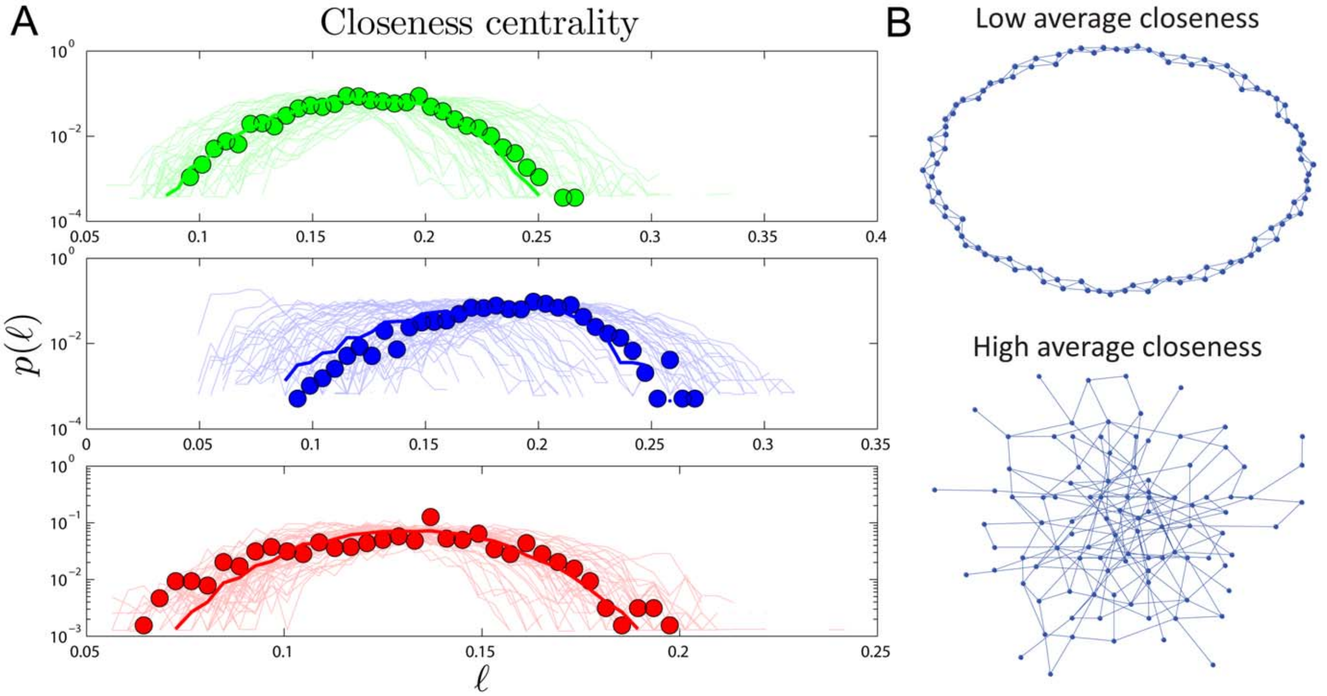

Figure 3. Closeness centrality. (A) Closeness (ℓ) distributions in human (green), yeast (blue), and fly (red). Heavy lines are the median values from 50 simulations, and light lines are results of individual simulations. (B) Examples of networks with low average closeness $\langle\ell\rangle = 0.06$ (top; each node is generally far away from most other nodes because there are no 'short cuts') and high average closeness $\langle\ell\rangle = 0.28$ (bottom; the random connections allow each node to be only a short distance from the other nodes). Note that both networks pictured here have the same number of nodes ($N = 100$) and roughly the same average degree (top: $\langle k\rangle = 4$, bottom: $\langle k\rangle = 3.7$).
doi:10.1371/journal.pone.0039052.g003

measured for the present-day network. Although our model accurately reproduces the present-day values of these quantities, there is no direct evidence that the simulated trajectories are correct; rather, these are predictions of our model. Figure 6 shows that both modularity Q and diameter D increase with time. These are not predictions that can be tested yet for biological systems, since there is no time-resolved data yet available for PPI evolution. Time-resolved data is only currently available for various social networks (links to websites, co-authorship networks, etc.). Interestingly, the diameters of social networks are found to shrink over time [74]. Our model predicts that PPI networks differ from these social networks in that their diameters grow over time. In addition to Q and D, we tracked the evolutionary trajectories of several other quantities: the evolution of the global clustering coefficient, the rate of signal propagation, the size of the largest connected component (Figure S3), as well as betweenness and degree values for individual nodes (Figure S4). See SI for details.

Discussion

The relevance of selection to PPI network evolution has been a topic of considerable debate [75], particularly in the context of higher-order network features, such as modularity. A number of authors have argued that specific selection programs are required to generate modular networks, such as oscillation between different evolutionary goals [76–81]. However, previous work has shown that gene duplication by itself, in the absence of both natural selection and neofunctionalization, can generate modular networks [82,83]. Consistent with the findings of [82,83], modularity in our model is primarily generated by gene duplications (Figure S11; see SI for sensitivity analysis). Unfortunately, duplication-only models err in their predictions of other network properties (Tables 2 and S2; Figure S12). A well-known problem with duplication models is that they generate excessively fragmented networks, with only about 20% of the proteins in the largest component. This is in sharp contrast to real PPI networks, which have 73% to 89% of their proteins in the largest component. Neofunctionalization-only models have most of their proteins in the largest component, but are significantly less modular than real networks. As shown in Table 2, by modeling duplication and neofunctionalization simultaneously, the DUNE model generates networks which have the modularity found in duplication-only models, while retaining most proteins in the largest component. This lends support to the idea that gene duplication contributes to the modularity found in real biological networks, and that protein modules can arise under neutral evolution, without requiring complicated assumptions about selective pressures. This is consistent with recent experimental work characterizing a real-world fitness landscape, showing that it is primarily shaped by neutral evolution [84].

Previous estimates of NE rates in eukaryotes have varied widely, generally falling in the range of 100 to 1000 changes/genome/Myr [24,46,85], or on the order of 0.1 change/gene/Myr. However, more recent empirical work has identified several problems with the methods used to obtain these estimates, suggesting that *de novo* link creation is much less common than previously thought [48]. This is consistent with our model. The best-fit values of our NE rate μ are in the range of 10^{-5} to 10^{-4}/gene/Myr (Table 1), which in all three organisms are considerably slower than the duplication rates d.

Biologically, many of the interactions created by our neofunctionalization mechanism are expected to initially be weak, non-functional interactions. The results of [86] suggest that strong

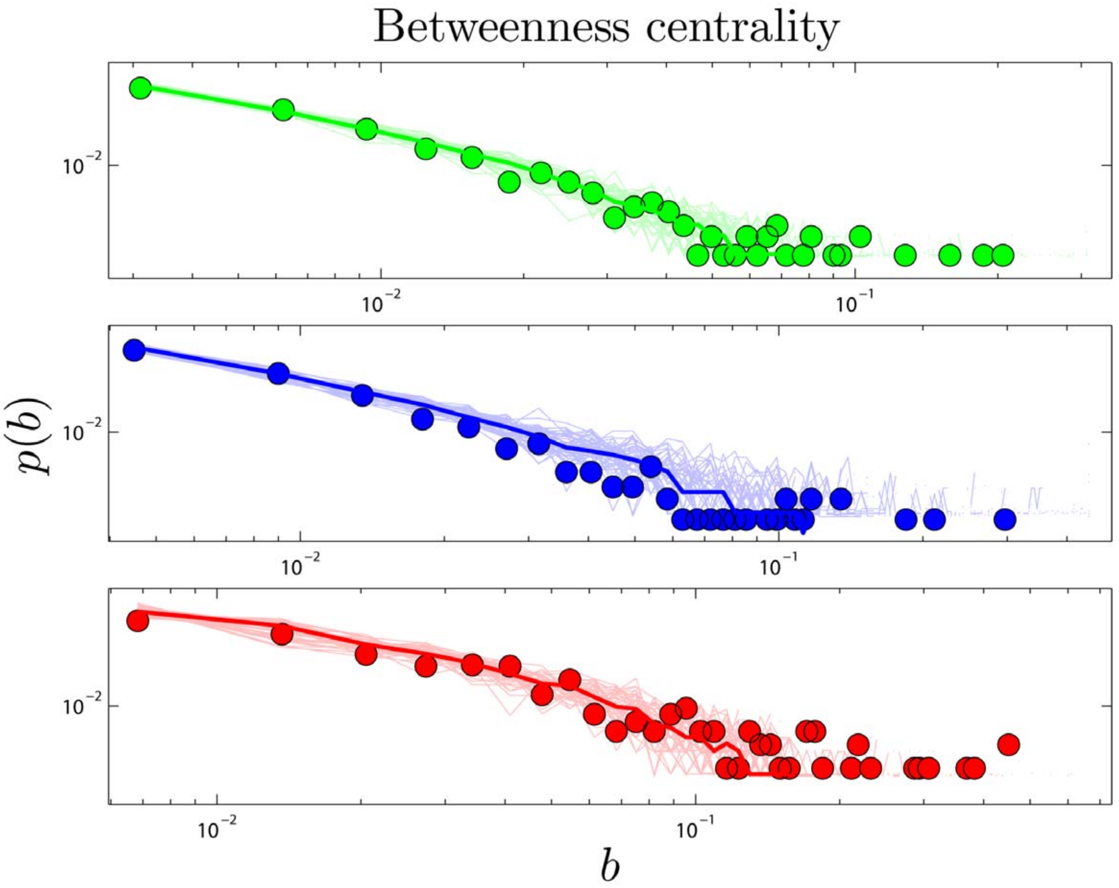

Figure 4. Betweenness centrality. Betweenness (*b*) distributions in human (green), yeast (blue), and fly (red). Heavy lines are the median values from 50 simulations, and light lines are results of individual simulations.
doi:10.1371/journal.pone.0039052.g004

functional interactions are correlated with hydrophobicity, which in turn is correlated with promiscuity. We posit that initially weak, non-functional interactions are an essential feature of PPI evolution, as they provide the 'raw material' for the subsequent evolution of functional interactions. If this reasoning is correct, one consequence should be that hub proteins are, on average, more important to the cell than non-hub proteins. This has been found to be true: both degree [54] and betweenness centrality [65] have positive correlations with essentiality, indicating that hub proteins are often critical to the cell's survival.

We have described here a model for how eukaryotic protein networks evolve. The model, called DUNE, implements two biological mechanisms: (1) gene duplications, leading to a superfluous copy of a protein that can change rapidly under new selective pressures, giving new relationships with other proteins and (2) a protein can undergo random mutations, leading to neofunctionalization, the *de novo* creation of new relationships with other proteins. Neofunctionalization can lead to assimilation, the formation of extra novel interactions with the other proteins in the target's neighborhood. Biological evidence suggests that this type of mechanism exists. Our specific implementation is based on a simple geometric surface-compatibility argument for the observed transitivity in PPI networks. This is, of course, a heavily simplified model of PPI network evolution, and there are many biological factors which have not been included. However, our relatively simple model shows good agreement with 10 topological properties in yeast, fruit flies, and humans. One finding is that PPI networks can evolve modular structures, just from these random forces, in the absence of specific selection pressures. We also find that the most central proteins also tend to be the oldest. This suggests that looking at the structures of present-day protein networks can give insight into their evolutionary history.

Methods

Genome-wide PPI screens have a high level of noise [19], and specific interactions correlate poorly between data sets [52]. We found that several large-scale features differed substantially between types of high-throughput experiments (see SI). Due to concerns about the accuracy and precision of data obtained through high-throughput screens, we chose to work with 'high-confidence' data sets consisting only of pairwise interactions confirmed in small-scale experiments, which we downloaded from the public HitPredict database [87]. We found sufficient high-confidence data in yeast (*S. cerevisiae*), fruit flies (*D. melanogaster*), and humans (*H. sapiens*).

All simulations and network feature calculations were carried out in Matlab. Our scripts are freely available for download at http://ppi.tinybike.net. We computed betweenness centralities, clustering coefficients, shortest paths, and component sizes using the MatlabBGL package. Modularity values were calculated with the algorithm of [88]. All comparisons (except the degree distribution) are between the largest connected components of the simulated and experimental data.

Due to the human network's somewhat larger size, most dynamical features were calculated once per 50 time steps for the

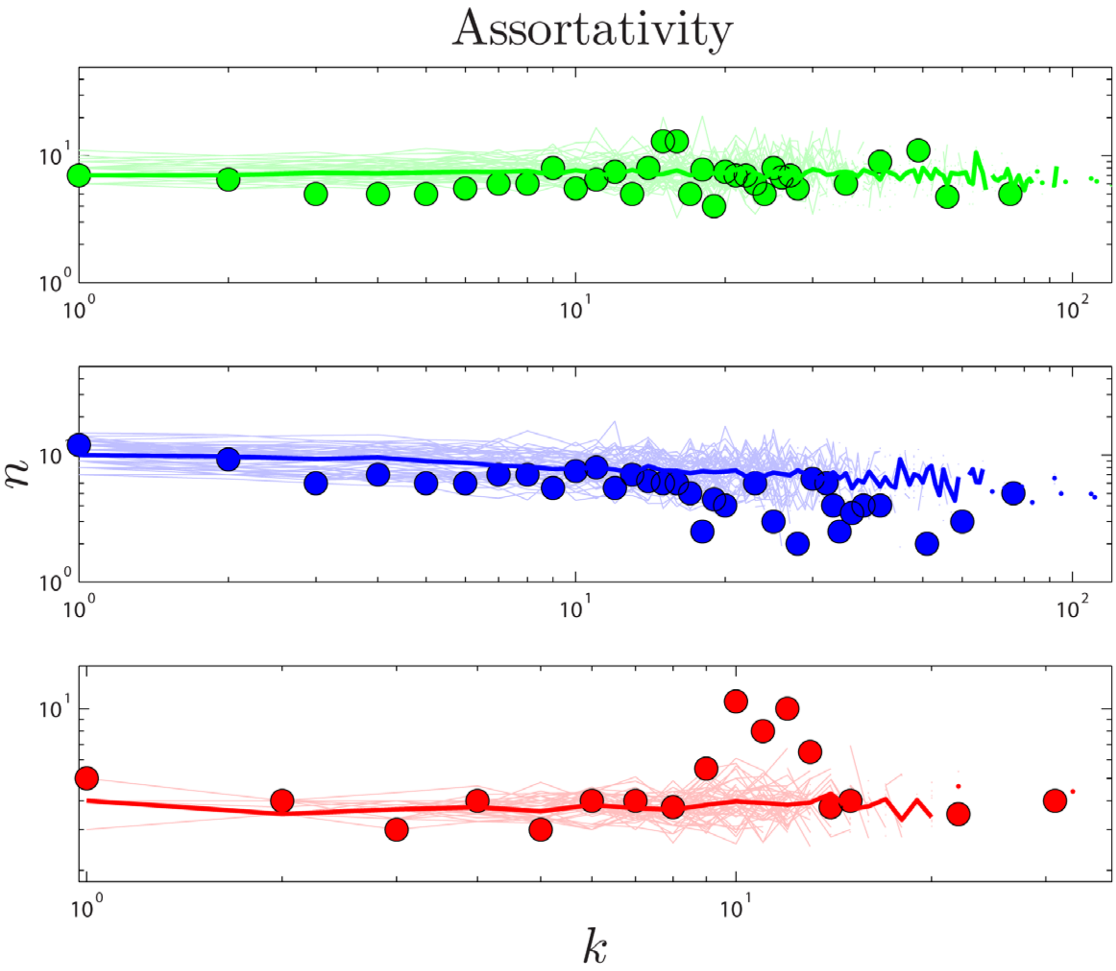

Figure 5. Assortativity. Median nearest-neighbor degree vs. degree in human (green), yeast (blue), and fly (red). Heavy lines are the median values from 50 simulations, and light lines are results of individual simulations.
doi:10.1371/journal.pone.0039052.g005

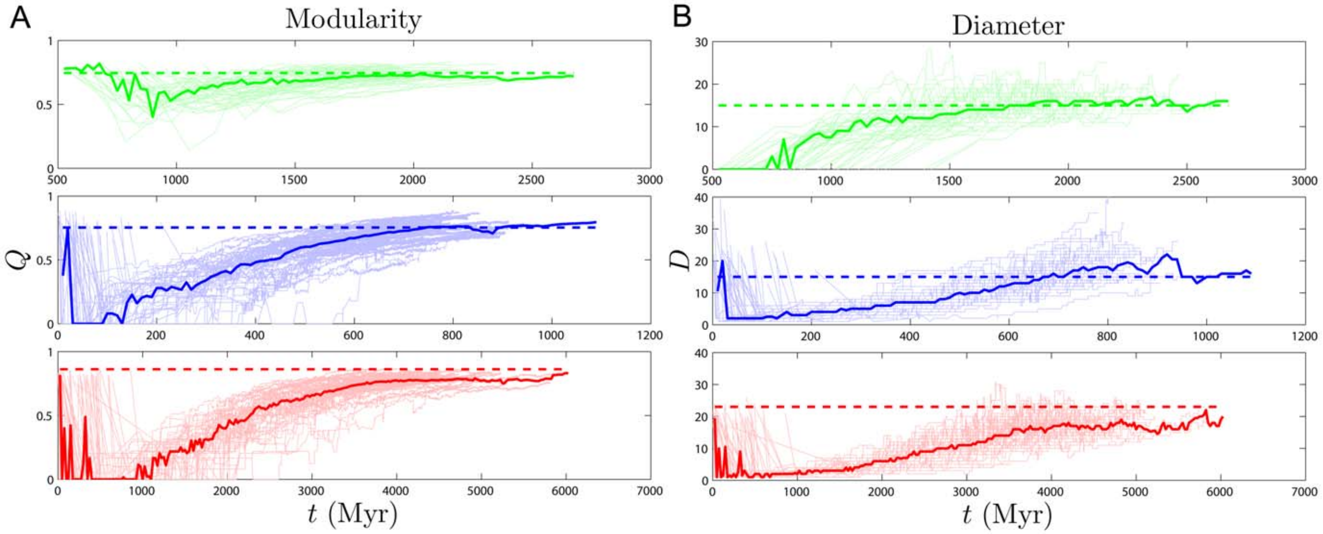

Figure 6. Modularity and diameter. (A) Modularity Q and (B) diameter D are both predicted to grow with time in human (green), yeast (blue), and fly (red). Light lines indicate the evolutionary trajectories of 50 individual simulations, and the heavy line is the median value. The modularity and diameter of the empirical data are shown as dashed horizontal lines. Time traces occasionally do not start at $t=0$ because these simulations spend the first few time steps in a completely disconnected state, so the dynamical quantities are undefined. (See Figure 9 for other dynamical plots.).
doi:10.1371/journal.pone.0039052.g006

human network, but were updated at every time step in the yeast and fly networks. For dynamical plots, the y coordinates of the trend line are medians-of-medians. The amount of time elapsed per time step (the x coordinate) varies between simulations. We binned the time coordinates to the nearest 10 million years for yeast and fly, and 25 million years for human. When multiple values from the same simulation fell within the same bin, we used the median value. We then calculated the median value between simulations. Scatter plot trend lines are calculated in a similar way. The trend line represents the median response variable (C, b, or ℓ) value over all nodes within a single simulation with degree k. The y coordinate of the trend line is therefore the median (across 50 simulations) of these median response variables. This median-of-medians includes all simulations that have nodes of a given degree.

Supporting Information

File S1 Supporting information text.
(PDF)

Figure S1 Hierarchical clustering. Median clustering coefficient vs. degree in human (green), yeast (blue), and fly (red). Heavy lines are the median values from 50 simulations, and light lines are results of individual simulations.
(TIF)

Figure S2 Older proteins are more central. Simulations of a protein's age index (time since introduction into the network) vs. degree (k), betweenness (b), and closeness (ℓ) centrality, for human (green), yeast (blue), and fly (red). The oldest proteins are on the *left* in this figure, and the proteins get younger moving to the right. There is an approximately monotonic increase in centrality with age.
(TIF)

Figure S3 Dynamical features. Shown are the evolution of (A) the largest component size, (B) the fraction of orphan proteins, (C) the global clustering coefficient, and (D) the second-largest eigenvalue of the walk matrix, in human (green), yeast (blue), and fly (red). Light lines indicate the evolutionary trajectories of 50 individual simulations, and the heavy line is the median value. Empirical data values are shown as a dashed line, where available.
(TIF)

Figure S4 Individual protein centrality scores. Evolution of degree (A) and betweenness (B) for proteins introduced to the network at different times in humans (top), yeast (middle), and flies (bottom). The 1st protein (one of the two initial proteins) is shown in red, the 6th protein in black, the 11th protein in blue, and the 101st protein in green. Curves are median values from 50 simulations.
(TIF)

Figure S5 Laplacian eigenvector participation. Elements of the eigenvector of the Laplacian matrix (defined as $\mathbf{K}-\mathbf{A}$, where $\mathbf{K}$ is a diagonal matrix with the degree of node i as element K_{ii}) associated with the largest eigenvalue vs. protein age index (time of introduction) in the yeast simulation. Details of this method are discussed in [73]. Heavy lines are the median values from 50 simulations, and light lines are results of individual simulations. The inset plot shows the trend line with a rescaled y-axis.
(TIF)

Figure S6 GO-slim profiles. Shown are profiles for the 100 oldest and 100 youngest proteins, as measured by S-value, in the yeast PPI network.
(TIF)

Figure S7 Walk matrix eigenvalues. Shown are eigenvalue (λ) distributions in human (green), yeast (blue), and fly (red). Heavy lines are the median values from 50 simulations, and light lines are results of individual simulations.
(TIF)

Figure S8 Error tolerance. Shown are error tolerance curves in human (green), yeast (blue), and fly (red). Circles indicate proteins deleted randomly, and squares indicate proteins deleted starting with the most well-connected protein and removing proteins in descending order.
(TIF)

Figure S9 Principal component analysis. Shown are the factor loadings and scores on the first two principal components. Data scores are shown in red, and blue lines represent feature loadings.
(TIF)

Figure S10 Betweenness vs. degree. Shown are median betweenness vs. degree values in human (green), yeast (blue), and fly (red). Heavy lines are the median values from 50 simulations, and light lines are results of individual simulations.
(TIF)

Figure S11 Sensitivity analysis. Heat maps represent median values for 10 simulations per parameter combination of the yeast network. Left: ϕ and a are varied, d and μ values are kept fixed. Right: d and μ varied, ϕ and a kept fixed.
(TIF)

Figure S12 Model comparison. Comparison of five other models to the yeast PPI network: Vázquez [29] (green), Berg [85] (red), random geometric [89] (dark blue), MpK desolvation [52] (purple), and ER random graph [90] (brown). For reference, DUNE model results are shown as a black line. Dots represent high-confidence experimental yeast data, and solid lines are median values over 50 simulations.
(TIF)

Table S1 Scaling exponents. Distributional exponents ($p(k)\sim k^{-\gamma}$, $p(b)\sim b^{-\beta}$) were estimated using the maximum likelihood method of [91]. Other exponents ($C\sim k^{-\xi}$, $b\sim k^{\alpha}$, $n\sim k^{-\delta}$) were estimated using nonlinear regression. Due to the relatively small sizes of the data sets, there is considerable uncertainty in these estimates.
(PDF)

Table S2 SMAPE values. Symmetric mean absolute percentage error (SMAPE) of simulation versus experiment in yeast (Eq. ??). 'E.T.' is the error tolerance curve with random protein removal, and 'E.T. (k)' is the error tolerance curve with highest-degree proteins removed first. 'DUNE' is the model described here, 'Vázquez' is the DU-only model of [29], 'Berg' is the link dynamics model [85], 'RG' is random geometric [89], 'MpK' is the physical desolvation model presented in [52], and 'ER' is an Erdös-Rényi random graph [90]. For each comparison, the lowest value is shown in bold.
(PDF)

Acknowledgments

We would like to thank Tom MacCarthy and Anna Panchenko for valuable discussions.

Author Contributions

Conceived and designed the experiments: GJP SP KAD. Performed the experiments: GJP. Analyzed the data: GJP KSP. Contributed reagents/materials/analysis tools: GJP SP KSP KAD. Wrote the paper: GJP KAD.

References

1. Hughes D (2003) Microbial genetics: Exploiting genomics, genetics and chemistry to combat antibiotic resistance. Nature Reviews Genetics 4: 432–441.
2. Cirz R, Chin J, Andes D, de Crécy-Lagard V, Craig W, et al. (2005) Inhibition of mutation and combating the evolution of antibiotic resistance. PLoS Biology 3: e176.
3. Taylor I, Linding R, Warde-Farley D, Liu Y, Pesquita C, et al. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nature Biotechnology 27: 199–204.
4. Lynch M, O'Hely M, Walsh B, Force A (2001) The probability of preservation of a newly arisen gene duplicate. Genetics 159: 1789–1804.
5. Ting CT, Tsaur SC, Sun S, Browne W, Chen YC, et al. (2004) Gene duplication and speciation in Drosophila: evidence from the Odysseus locus. Proc Natl Acad Sci USA 101: 12232–12235.
6. Dutkowski J, Tiuryn J (2009) Phylogeny-guided interaction mapping in seven eukaryotes. BMC Bioinformatics 10: 393.
7. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, et al. (2000) Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc Natl Acad Sci USA 97: 1143–1147.
8. Uetz P, Giot L, Cagney G, Mansfield T, Judson R, et al. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403: 623–627.
9. Krogan N, Cagney G, Yu H, Zhong G, Guo X, et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440: 637–643.
10. Yu H, Braun P, Yildirim M, Lemmens I, Venkatesan K, et al. (2008) High-quality binary protein interaction map of the yeast interactome network. Science 5898: 104–110.
11. Marcotte E, Pellegrini M, Ng H, Rice D, Yeates T, et al. (1999) Detecting protein function and protein-protein interactions from genome sequences. Nature 402: 86–90.
12. Pellegrini M, Marcotte E, Thompson M, Eisenberg D, Yeates T (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 96: 4285–4288.
13. Valencia A, Pazos F (2002) Computational methods for the prediction of protein interactions. Current Opinion in Structural Biology 12: 368–373.
14. Gomez S, Noble W, Rzhetsky A (2003) Learning to predict protein-protein interactions from protein sequences. Bioinformatics 19: 1875–1881.
15. Jothi R, Kann M, Przytycka T (2005) Predicting protein-protein interaction by searching evolutionary tree automorphism space. Bioinformatics 21: 241–250.
16. Liu Y, Liu N, Zhao H (2005) Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. Bioinformatics 21: 3279–3285.
17. Shoemaker B, Panchenko A (2007) Deciphering protein-protein interactions. part II. computational methods to predict protein and domain interaction partners. PLoS Computational Biology 3: e43.
18. Burger L, van Nimwegen E (2008) Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Mol Syst Biol 4: 165.
19. Deane C, Salwiński L, Xenarios I, Eisenberg D (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. Molecular & Cellular Proteomics 1: 349–356.
20. Yamada T, Bork P (2009) Evolution of biomolecular networks – lessons from metabolic and protein interactions. Nature Reviews Molecular Cell Biology 10: 791–803.
21. Ohno S (1970) Evolution by Gene Duplication. Springer-Verlag, New York.
22. Zhang J (2003) Evolution by gene duplication: an update. Trends in Ecology & Evolution 18: 292–298.
23. Xiao H, Jiang N, Schaffner E, Stockinger E, van der Knaap E (2008) A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science 319: 1527–1530.
24. Wagner A (2003) How the global structure of protein interaction networks evolves. Proc R Soc Lond B 270: 457–466.
25. Koch A (1972) Enzyme evolution. i. the importance of untranslatable intermediates. Genetics 72: 297–316.
26. Taylor J, Raes J (2004) Duplication and divergence: the evolution of new genes and old ideas. Annual Review of Genetics 38: 615–643.
27. Lynch M, Conery J (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.
28. Maslov S, Sneppen K, Eriksen K KA amd Yan (2004) Upstream plasticity and downstream robustness in evolution of molecular networks. BMC Evolutionary Biology 4: 9.
29. Vázquez A, Flammini A, Maritan A, Vespignani A (2003) Modeling of protein interaction networks. ComPlexUs 1: 38–44.
30. Kellis M, Birren B, Lander E (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 428: 617–624.
31. Gu X, Zhang Z, Huang W (2005) Rapid evolution of expression and regulatory divergences after yeast gene duplication. Proc Natl Acad Sci USA 102: 707–712.
32. Lynch M, Walsh B (1998) Genetics and Analysis of Quantitative Traits. Sinauer, Sunderland, MA.
33. Ispolatov I, Yuryev A, Mazo I, Maslov S (2005) Binding properties and evolution of homodimers in protein-protein interaction networks. Nucleic Acids Res 33: 3629–3635.
34. Jones S, Thornton J (1996) Principles of protein-protein interactions. Proc Natl Acad Sci USA 93: 13–20.
35. Tovchigrechko A, Vakser I (2001) How common is the funnel-like energy landscape in protein-protein interactions? Protein Science 10: 1572–1583.
36. Wu F, Towfic F, Dobbs D, Honavar V (2007) Analysis of protein-protein dimeric interfaces. In: IEEE International Conference on Bioinformatics and Biomedicine. Fremont, CA.
37. Vinogradov A, Anatskaya O (2009) Loss of protein interactions and regulatory divergence in yeast whole-genome duplicates. Genomics 93: 534–542.
38. Escriva H, Bertrand S, Germain P, Robinson-Rechavi M, Umbhauer M, et al. (2006) Neofunction-alization in vertebrates: The example of retinoic acid receptors. PLoS Genetics 2: e102.
39. Gillespie D (1977) Exact stochastic simulation of coupled chemical reactions. Journal of Physical Chemistry 81: 2340–2361.
40. Gu Z, Cavalcanti A, Chen F, Bouman P, Li W (2002) Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. Molecular Biology and Evolution 19: 256–262.
41. Gao L, Innan H (2004) Very low gene duplication rate in the yeast genome. Science 306: 1367–1370.
42. Lynch M, Conery J (2003) The evolutionary demography of duplicate genes. Journal of Structural and Functional Genomics 3: 35–44.
43. Osada N, Innan H (2008) Duplication and gene conversion in the Drosophila melanogaster genome. PLoS Genetics 4: e1000305.
44. Cotton J, Page R (2005) Rates and patterns of gene duplication and loss in the human genome. Proceedings of the Royal Society B 272.
45. Pan D, Zhang L (2007) Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: a novel strategy to estimate gene duplication rates. Genome Biol 8: R158.
46. Beltrao P, Serrano L (2007) Specificity and evolvability in eukaryotic protein interaction networks. PLoS Computational Biology 3.
47. Lynch M, Sung W, Morris K, Coffey N, Landry C, et al. (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci USA 105: 9272–9277.
48. Gibson T, Goldberg D (2009) Questioning the ubiquity of neofunctionalization. PLoS Computational Biology 5: e1000252.
49. Kimura M (1957) Some problems of stochastic processes in genetics. Ann Math Stat 28: 882–901.
50. Walsh J (1995) How often do duplicated genes evolve new functions? Genetics 139: 421–428.
51. He X, Zhang J (2005) Rapid subfunctionalization accompanied by prolonged and substantial neo-functionalization in duplicate gene evolution. Genetics 169: 1157–1164.
52. Deeds E, Ashenberg O, Shakhnovich E (2006) A simple physical model for scaling in protein-protein interaction networks. Proc Natl Acad Sci USA 103: 311–316.
53. Patil A, Kinoshita K, Nakamura H (2010) Domain distribution and intrinsic disorder in hubs in the human protein-protein interaction network. Protein Science 19: 1461–1468.
54. Jeong H, Mason S, Barabási A, Oltvai Z (2001) Lethality and centrality in protein networks. Nature 411: 41–42.
55. Ku H, Vision T, Liu J, Tanksley S (2000) Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc Natl Acad Sci USA 97: 9121–9126.
56. Travers J, Milgram S (1969) An experimental study of the small world problem. Sociometry 32: 425–443.
57. Leskovec J, Horvitz E (2008) In:Proceedings of the 17th international conference on World Wide Web. ACM, New York.
58. Yook S, Oltvai Z, Barabási A (2004) Functional and topological characterization of protein interaction networks. Proteomics 4: 928–942.
59. Newman M, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69: 026113.
60. Newman M (2006) Modularity and community structure in networks. Proc Natl Acad Sci USA 103: 8577–8582.
61. Maslov S, Sneppen K (2002) Specificity and Stability in Topology of Protein Networks. Science 296: 910–913.
62. von Mering C, Krause R, Snel B, Cornell M, Oliver S, et al. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 31: 399–403.
63. Ravasz E, Somera A, Mongru D, Oltvai Z, Barabási A (2002) Hierarchical organization of modularity in metabolic networks. Science 297: 1551–1555.
64. Barabási A, Oltvai Z (2004) Network biology: understanding the cell's functional organization. Nature Reviews Genetics 5: 101–113.
65. Joy M, Brock A, Ingber D, Huang S (2005) High-betweenness proteins in the yeast protein interaction network. Journal of Biomedicine and Biotechnology 2: 96–103.
66. Zhao D, Liu Z, Wang J (2007) Duplication: a mechanism producing disassortative mixing networks in biology. Chin Phys Lett 24: 2766.

67. Wan X, Cai S, Zhou J, Liu Z (2010) Emergence of modularity and disassortativity in protein-protein interaction networks. Chaos 20: 045113.
68. Aloy P, Russell R (2002) Potential artefacts in protein-interaction networks. FEBS Lett 530: 253–254.
69. Hakes L, Pinney J, Robertson D, Lovell S (2008) Protein-protein interaction networks and biology – what's the connection? Nature Biotechnol 26: 69–72.
70. Woese C (1987) Bacterial evolution. Microbiol Rev 51: 221–271.
71. Krapivsky P, Redner S (2001) Organization of growing random networks. Physical Review E 63: 066123.
72. Qin H, Lu H, Wu W, Li W (2003) Evolution of the yeast protein interaction network.Proc Natl Acad Sci USA 100: 12820–12824.
73. Zhu G, Yang H, Yang R, Ren J, Li B, et al. (2012) Uncovering evolutionary ages of nodes in complex networks. Eur Phys Jour B 85: 106.
74. Leskovec J, Kleinberg J, Faloutsos C (2005) Graphs over time: densification laws, shrinking diameters and possible explanations. Knowledge Discovery in Databases: PKDD 2005.
75. Lynch M (2007) The evolution of genetic networks by non-adaptive processes. Nat Rev Genet 8: 803–813.
76. Lipson H, Pollack J, Suh N (2002) On the origin of modular variation. Evolution 56: 1549–1556.
77. Kashtan N, Alon U (2005) Spontaneous evolution of modularity and network motifs. Proc Natl Acad Sci USA 102: 13773–13778.
78. Komurov K, White M (2007) Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. Molecular Systems Biology 3: 110.
79. Wagner G, Pavlicev M, Cheverud J (2007) The road to modularity. Nature Reviews Genetics 8: 921–931.
80. Espinosa-Soto C, Wagner A (2010) Specialization can drive the evolution of modularity. PLoS Computational Biology 6: e1000719.
81. Soyer O (2010) Fate of a duplicate in a network context. In: Dittmar K, Liberles D, editors, Evolution After Gene Duplication, Wiley-Blackwell. 215–228.
82. Hallinan J (2004) Gene duplication and hierarchical modularity in intracellular interaction networks. BioSystems 74: 51–62.
83. Solé R, Valverde S (2008) Spontaneous emergence of modularity in cellular networks. J R Soc Interface 5: 129–133.
84. Hietpas R, Jensen J, Bolon D (2011) Experimental evolution of a fitness landscape. Proc Natl Acad Sci USA 108: 7896–7901.
85. Berg J, Lassig M, Wagner A (2004) Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. BMC Evolutionary Biology 4: 51–63.
86. Heo M, Maslov S, Shakhnovich E (2011) Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. Proc Natl Acad Sci USA 108: 4258–4263.
87. Patil A, Nakai K, Nakamura H (2011) Hitpredict: a database of quality assessed protein-protein interactions in nine species. Nucleic Acids Research 39 (suppl 1): D744–D749.
88. Blondel V, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008: P10008+.
89. Pržulj N, Corneil D, Jurisica I (2004) Modeling interactome: scale-free or geometric? Bioinformatics 20: 3508–3515.
90. Erdős P, Rényi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5: 17–61.
91. Clauset A, Shalizi C, Newman M (2009) Power-law distributions in empirical data. SIAM Review 51: 661–703.