BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 3 2011, pages 376–382
doi:10.1093/bioinformatics/btq623

Systems biology

Advance Access publication November 9, 2010

Improving evolutionary models of protein interaction networks

Todd A. Gibson and Debra S. Goldberg*

Department of Computer Science, University of Colorado, 430 UCB, Boulder, CO 80309, USA

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Theoretical models of biological networks are valuable tools in evolutionary inference. Theoretical models based on gene duplication and divergence provide biologically plausible evolutionary mechanics. Similarities found between empirical networks and their theoretically generated counterpart are considered evidence of the role modeled mechanics play in biological evolution. However, the method by which these models are parameterized can lead to questions about the validity of the inferences. Selecting parameter values in order to produce a particular topological value obfuscates the possibility that the model may produce a similar topology for a large range of parameter values. Alternately, a model may produce a large range of topologies, allowing (incorrect) parameter values to produce a valid topology from an otherwise flawed model. In order to lend biological credence to the modeled evolutionary mechanics, parameter values should be derived from the empirical data. Furthermore, recent work indicates that the timing and fate of gene duplications are critical to proper derivation of these parameters.

Results: We present a methodology for deriving evolutionary rates from empirical data that is used to parameterize duplication and divergence models of protein interaction network evolution. Our method avoids shortcomings of previous methods, which failed to consider the effect of subsequent duplications. From our parameter values, we find that concurrent and existing existing duplication and divergence models are insufficient for modeling protein interaction network evolution. We introduce a model enhancement based on heritable interaction sites on the surface of a protein and find that it more closely reflects the high clustering found in the empirical network.

Contact: Debra@Colorado.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 22, 2010; revised on September 27, 2010; accepted on November 3, 2010

1 INTRODUCTION

The study of biological networks is a promising area for arriving at evolutionary inference. Gene duplication is regarded as the primary evolutionary phenomenon driving protein network growth. In theoretical models, gene duplication is represented through the random copying of one of the proteins (nodes) in the network. Each duplication is accompanied by 'link dynamics', the gain and loss of interactions in the network based on theories of gene duplicate fates. Model parameters control the probability at which interactions are gained and lost with each duplication. Two models in particular, Solé *et al.* (2002) and Vázquez *et al.* (2003) are exemplars of this type of theoretical model.

The strategy used to affirm the evolutionary mechanics captured by these models is to compare the topology generated by these networks with empirical biological networks. The models are seen to be validated when the theoretical and empirical networks share similar topological characteristics.

It therefore follows that proper parameterization of the models is paramount to generating a network, which is plausible given the model mechanics. Vázquez *et al.* (2003) used the average number of empirical protein interactions from empirical data (average degree) as a constraint to 'tune' model parameter values, which produced a similar theoretical network. Solé *et al.* (2002) derived parameter estimates directly from empirical measurements based on the fact that the model parameters represent the gain and loss of interactions after gene duplication. Their interaction gain and loss rate parameter values were based on examination of interactions of known duplicates (i.e. paralogs; Wagner, 2001), the average degree and mathematical constraints.

The mechanics featured in the Vázquez and Solé models are essential to this study. Gene duplication, loss of redundant interactions between duplicates, the formation of new interactions and interactions which form between paralogs when a self-interacting protein's gene is duplicated are all evolutionary phenomena which have direct analogs in the Vázquez and Solé models. It is these evolutionarily plausible mechanics to which we give special attention. Can these mechanics be enhanced to improve model performance while preserving the biological plausibility of the enhancements?

Aside from Vázquez and Solé, there are numerous models that have furthered our understanding of protein interaction network evolution. One notable example is Berg *et al.* (2004), who construct a stochastic model based on empirical observations. The model features protein interaction gain and loss, but the interaction loss rate is predicated on preserving a constant network connectivity rather than on a biologically identifiable phenomena. In another example, Pržulj *et al.* (2004) find that a geometric random graph has a good topological fit to empirical protein interaction networks. Their geometric random graph is constructed by connecting nodes based on their proximity from each other when arranged randomly on a two-dimensional metric space. The mechanics of the model are not evolutionarily plausible; they do not help us elucidate the evolutionary processes that have contributed to the formation of the network topology.

A strength of the Vázquez model over other models with biologically faithful mechanics is that it does not require

*To whom correspondence should be addressed.

neofunctionalization, the *de novo* addition of interactions to the network. Recently, it has been revealed that neofunctionalization in protein interaction network evolution has not been shown to be prevalent as previously believed (Gibson and Goldberg, 2009a). The Solé *et al.* model features the prolific gain of *de novo* interactions. Likewise Berg *et al.* (2004) suggest a model with very fluid link dynamics, which is directly premised on ubiquitous *de novo* interaction gain (Gibson and Goldberg, 2009a; Wagner, 2003).

Given the drawbacks of the other models, Vázquez *et al.* stands as the best theoretical model, which promotes gene duplication and biologically plausible link dynamics as sufficient evolutionary mechanics for generating protein interaction networks. But despite the model's foundational prominence (Newman *et al.*, 2006), the parameterization used by Vázquez *et al.* (2003) merits a second look. As we show later, a wide variety of parameter values produce similar topological characteristics. This compromises the efficacy of the model's evolutionary mechanics. Selecting parameter values in order to produce a desired topology makes it problematic to attribute the desired topology to the evolutionary mechanics. The validation question should be 'Do *biologically justified* parameter values determined independently of the model produce topological characteristics similar to empirical?' The method of parameterization used in the published model requires further analysis to reaffirm the role of gene duplication and biologically plausible link dynamics play in the formation of protein interaction networks.

Here, we develop an improved methodology for measuring the rate of interaction gain and loss from empirical data. We assign these rates to the Vázquez *et al.* model parameters and identify the limits of the Vázquez *et al.* model. In an attempt to overcome identified limits of this model, we incorporate and assess two evolutionarily plausible modifications: subfunctionalization asymmetry and heritable homomeric interactions.

We then introduce a new evolutionary model featuring heritable interaction sites. That is, we associate a protein's interactions with one or more physical sites on the protein responsible for the interaction binding. As the model evolves, these sites are included in gene duplication events. Subsequent to duplication, subfunctionalization of interactions is applied through the interaction sites rather than to each interaction independently. We find that the model is more effective than the Vázquez *et al.* model at reproducing clustering found in the empirical network, suggesting the importance of heritable interaction sites in the evolution of protein interaction networks.

2 METHODS

2.1 Network measures of subfunctionalization and heteromerization

The Vázquez *et al.* model parameters reflect evolutionary rates that can be estimated from extant data. (The Vázquez *et al.* model and the empirical dataset are described in the Supplementary Materials.) The subfunctionalization rate, that is the proportion of redundant interactions lost between gene duplicates, can be measured as the number of interacting neighbors not shared by both gene duplicates versus the total number of neighbors the duplicates interact with. Previous studies have calculated proportions of redundant interaction retention or loss by simple tallies of neighbors in the empirical network (Beltrao and Serrano, 2007; Berg *et al.*, 2004; Chung *et al.*, 2006; He and Zhang, 2005; Wagner, 2001, 2002, 2003). This method fails to account for duplications that occur subsequent to the duplication being measured (Gibson and Goldberg, 2009a). We have abandoned the previous method in favor of utilizing interaction data drawn from an evolutionary framework presented in Gibson and Goldberg (2009b). Briefly, a phylogeny of yeast species and phylogenetic gene trees were combined with interaction data to reconstruct a putative timing of duplication events along with associated interaction gain and loss. Further details are available in the Methods of Supplementary Material. In Section 4.1, we present an argument for our alternative rate estimation methodology and identify issues with the earlier methodology.

As described in Gibson and Goldberg (2009b), the timed duplications are binned into phylogenetic time periods. The subfunctionalization rate is calculated at each phylogenetic time period. In the absence of more detailed timing data, duplication events are assumed to have occurred simultaneously within each phylogenetic time period. That is, the subfunctionalization rate for a phylogenetic time period is measured only after all duplications (and corresponding redundant interaction losses) within that time period have been computed. This is semantically equivalent to treating all duplications within a phylogenetic time period as the product of a segmental- or whole-genome duplication event.

More specifically, the subfunctionalization rate is calculated as:

$$\frac{\sum_t l_t}{\sum_t p_t}$$

where l_t is the number of interactions at phylogenetic time period t lost due to subfunctionalization, and p_t is the number of potential interactions added after all duplications for time period t but prior to interaction losses due to subfunctionalization. Additional detail is provided in the Supplementary Materials.

Supplementary Figure 1A shows the subfunctionalization rates measured in the evolutionary framework. The mean subfunctionalization rate is 0.887. That is, 88.7% of redundant interactions are lost between a neighbor and either the progeny or progenitor.

Like subfunctionalization, the proportion of lost paralogous interactions (interactions between homomeric progeny and progenitor duplicates which are lost) is calculated using a similar approach. According to the evolutionary framework (Gibson and Goldberg, 2009b), the ancestral protein is assumed to have been self-interacting if either the progenitor or progeny protein is self-interacting, or if an interaction exists between them. Including interactions between the paralogs mitigates the paucity of self-interactions reported by some assays (Gibson and Goldberg, 2009a).

The measure for the entire evolutionary history is aggregated similarly to the subfunctionalization method and is presented in the Methods of Supplementary Material. Using this measure, the probability of losing an interaction between self-interacting progeny and progenitor proteins is 0.539 (Supplementary Fig. 1B). The complement, that is, the probability of an interaction between paralogs being retained is $1-0.539=0.461$. Note that this is subtly different from the original Vázquez *et al.* model parameter. Our calculated parameter is the paralogous interaction retention rate among homomeric duplications. The Vázquez *et al.* parameter is the combined likelihood that the progenitor protein is self-interacting and the paralogous interaction survives. The computed Vázquez rate is 0.24 (see Methods of Supplementary Material). We also estimated parameters for several alternative yeast datasets. For various datasets, the subfunctionalization rate is relatively stable, but the paralogous interaction retention rate varies with connectivity. Full details can be found in the Supplementary Materials and are summarized in Supplementary Table 6.

From our calculations of the subfunctionalization and paralogous heteromerization rates, we are able to derive a general rate of interaction change for the species. Our derived rate is 1.2×10^{-5}changed interactions/My. This comports well with a previously published estimate of 2.4×10^{-5}changed interactions/My (Beltrao and Serrano, 2007). Full details are provided in the Supplementary Materials.

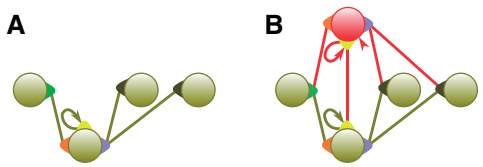

Fig. 1. The iSite model. (**A**) A simple network showing that protein interactions are tied to interaction sites on the protein surface. (**B**) A duplication event duplicates both the interactions and their iSites. The self-interacting iSite produces a self-interaction in the progeny and a paralogous interaction between the progenitor's and progeny's self-interacting iSite. Subfunctionalization is applied against all interactions of a redundant iSite (here paired by yellow, purple, orange, dark olive and green colors).

2.2 The iSite model

The iSite model integrates additional biologically plausible features while minimizing the increase in model complexity. The primary modification is to interaction semantics.

In the original Vázquez *et al.* model, each protein's interaction is independent from every other interaction the protein engages in. This independence is reflected in the implementation of subfunctionalization. The loss of one interaction from each pair of redundant interactions occurs independently of all other redundant pairs. We modify this by associating each interaction with an interaction site (iSite). An iSite is analogous to the physical contact one protein makes with another to form an interaction. Each interaction between two proteins requires two interaction sites, one on each protein. Figure 1 illustrates the iSite concept. There is a one-to-many relationship between iSites and interactions. Each iSite may be associated with many interactions, but each interaction involves only two iSites, one iSite on each protein participating in the interaction. The number of interactions associated with an iSite changes during evolution as shown on the neighbors in Figure 1B. Under this new model, the unit of redundancy subject to subfunctionalization is the iSite, not the interaction. For example, assume a protein has five interactions associated with a single iSite. Upon duplication, all five interactions of the iSite are subject to subfunctionalization as a single unit in either the progeny or progenitor protein. That is, asymmetry drives whether the progeny or progenitor iSite is selected, and then the subfunctionalization probability is applied to all interactions of the iSite.

This reconceptualization of the relationship between proteins and interactions also leads to a more biologically plausible implementation of homomeric duplication. The Vázquez *et al.* model does not directly incorporate self-interacting proteins. Instead, a parameter (p) is used as a probability for adding an interaction between the progeny and progenitor proteins. This probability corresponds to the combined likelihood that the progenitor protein is self-interacting and the paralogous interaction survives. The iSite model allows self interactions to be directly represented in the network. The iSite model repurposes p to reflect the probability that a self-interacting iSite is preserved. Conversely, $1-p$ is the probability an iSite is silenced in one of the duplicates due to subfunctionalization.

The mechanics of interaction loss are identical for both iSites featuring self interactions and other iSites. One iSite from each redundant progeny/progenitor iSite pair is selected with equal probability. Once an iSite is selected, each interaction within the iSite is silenced with probability q. Because all interactions within each redundant iSite are subject to silencing, redundant interaction pairs are not lost independently as in the Vázquez *et al.* model. Biologically, this acknowledges that individual mutations at an interaction site do not necessarily silence all protein interactions at that site.

2.3 Heritable homomers

The Vázquez *et al.* model obviates the need to include self-interacting proteins. (The Vázquez *et al.* model is described in the Supplementary Materials.) The parameter p serves as a surrogate, indicating the probability of adding an interaction between paralogs after each duplication. However, self-interactions and their heritability can be readily accommodated with few modifications to the model. Upon duplication of a self-interacting protein, two interactions are added: a self-interaction for the progeny and an interaction between the progenitor and progeny. Following the principle of subfunctionalization, two of the three interactions (two self-interactions plus the paralog-joining interaction) are redundant and may be lost according to the p parameter.

We enhanced the Vázquez *et al.* model with heritable homomers, with the parameter, p, repurposed to reflect the probability that the paralogous interaction and/or one of the redundant self-interactions would be lost due to subfunctionalization.

2.4 Asymmetry

The asymmetric divergence of gene duplicates can be readily modeled in protein interaction evolution. Asymmetric divergence has been observed in sequence mutations between gene duplicates (Byrne and Wolfe, 2007; Kellis *et al.*, 2004) as well as in coexpression neighbors and genetic interactions (Chung *et al.*, 2006; Wagner, 2002). The analog to sequence asymmetry in protein interaction networks is subfunctionalization asymmetry, the asymmetric loss of redundant interactions between gene duplicates (Wagner, 2002).

Both the Solé *et al.* (2002) and Vázquez *et al.* (2003) models feature subfunctionalization, but each handles asymmetry differently (Supplementary Fig. 5). Subfunctionalization in the Solé *et al.* model is 100% asymmetric: any redundant interaction loss after duplication is lost from the progeny protein. The Vázquez *et al.* model employs symmetric subfunctionalization: any lost redundant interactions are selected from either the progenitor or progeny protein with equal probability.

3 RESULTS

The topological properties of both the empirically derived yeast network and other networks can be found in Table 1. The top row of the table covers the empirical dataset. Additional datasets can be found in the Supplementary Materials. The bottom half of the table contains the topological properties of various theoretical models that are described later in the text. The 'Order' and 'Size' columns refer to the number of proteins and interactions, respectively. 'Triangles' and 'Connected Triples' refer to the components of the clustering coefficient. The clustering coefficient is defined as $C = \frac{3T}{\Gamma}$, where T is the number of triangles and Γ is the number of connected triples. A triangle is three fully connected nodes and a connected triple is a node connected to an unordered pair of other nodes (i.e. a path of length 2) (Newman, 2001). The clustering coefficient is the primary measure we use to scrutinize these models. It is a uniquely appropriate measure because it is sensitive to the model mechanics we study. The gain of interactions with each duplication, the loss of interactions as interactions are subfunctionalized, and the paralogous interactions formed between duplicates each have a measurable effect on the clustering coefficient (Gibson and Goldberg, 2009a). Additional topological measures can be found in the Supplementary Materials.

3.1 Clustering landscape

In order to identify the full range of clustering coefficients the Vázquez *et al.* model can generate, we constructed a heatmap from networks generated from a wide range of parameter values (Fig. 2). Full details of its construction are in the Methods of Supplementary Material. Figure 2 shows the clustering coefficients which are

Table 1. *Saccharomyces cerevisiae* topology

Dataset	Order	Size	Triangles	Connected triples	C	C of R.E.	Deg. distance difference	Homomeric proteins
Empirical	2647	5449	3299	66799	0.148	1.7×10^{-2}	0	412
Vázquez *et al.*	2647	1845 (34%)	44 (1%)	1477 (2%)	0.089	0.0	922	N/A
Heritable Vázquez	2647	2321 (42%)	22 (1%)	1279 (2%)	0.051	0.0	1280	561 (136%)
Asym-82% Vázquez	2647	2338 (42%)	117 (3%)	4021 (6%)	0.087	7.5×10^{-4}	552	N/A
Asym-91% Vázquez	2647	2806 (51%)	224 (6%)	7899 (11%)	0.085	1.9×10^{-3}	357	N/A
iSite model	2647	5515 (101%)	1699 (52%)	33088 (50%)	0.154	4.8×10^{-3}	334	652 (158%)

Topological properties related to the clustering coefficient are shown for the empirical data described in the Supplementary Materials. The large component of the combined data are also shown. The bottom rows feature mean values for 1000 runs of four models of protein interaction network evolution. The *Asym-* rows are the modified Vázquez *et al.* model at 82 and 91% asymmetric subfunctionalization. Percentages in parentheses are the proportion of the *empirical* dataset for the given value. *C of R.E* is the clustering coefficient of each network's random equivalent (see Methods of Supplementary Material). The *Size* (number of interactions) provides a comparative measure of the average degree since all networks in the table are of the same order. The degree distribution difference is the square root of the sum of squared degree difference between the empirical dataset and each of the theoretical models. A plot of the network degree distributions can be found in Supplementary Figure 3.

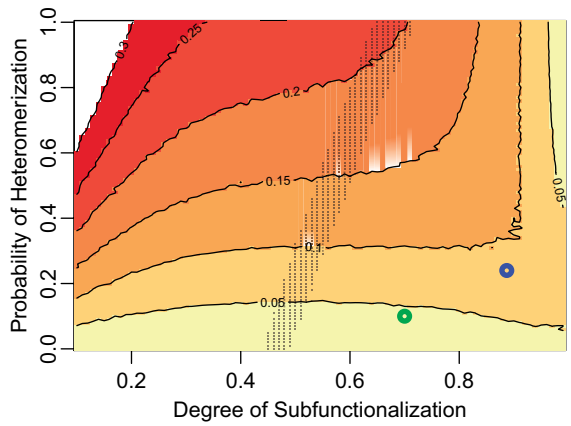

Fig. 2. The clustering landscape of Vázquez *et al.* Each data point is the mean clustering coefficient for 100 runs of the model at parameter values along the corresponding axes. Each network was constructed to 2647 nodes, the number of nodes in the empirical data (Table 1). The green point identifies the parameter values published in Vázquez *et al.* (2003) (see Supplementary Materials). The blue point identifies the parameter values derived from our analysis of the empirical data. The gray band covers those parameter values which generate network sizes (i.e. number of interactions) within 10% of 5449, the size of the empirical network.

possible using the Vázquez *et al.* model. The figure's 0.15 contour line is a close approximation to the 0.148 clustering coefficient of the empirical dataset (Table 1). The contour line indicates that a wide range of parameter values achieve the empirical clustering.

3.2 Vázquez *et al.* model and variants

Vázquez *et al.* (2003) assigned parameter values $q = 0.7$ (subfunctionalization) and $p = 0.1$ (heteromerization) to their model (see Supplementary Materials). These parameter values produce a clustering coefficient of 0.039 (green point in Fig. 2). When the Vázquez *et al.* model is initialized to the measures derived in Section 2.1 ($p = 0.24, q = 0.887,$ order $= 2647$), the model generates networks with a mean clustering coefficient of 0.108. The blue point in Figure 2 identifies where this result falls in the clustering landscape of the model. The model's clustering coefficient is distinctly lower than the empirical network clustering coefficient of 0.148 (see Table 1).

Protein interaction networks have high clustering coefficients relative to random equivalent networks (Goldberg and Roth, 2003; Solé *et al.*, 2002; Wagner, 2001). A necessary criterion of equivalency is that the random network maintain the same size (i.e. number of interactions) as the empirical network under study. In the absence of this equivalency restriction, a randomly generated network could engineer any clustering coefficient by simply varying its size. By extension, models of network evolution boasting high clustering coefficients should also be evaluated on the size of the generated network.

Using the empirically derived parameter estimates, the Vázquez *et al.* model generates a mean size of 1845. This is only 34% of the 5449 interactions in the empirical network. This low number is also reflected in the triangles and triples which comprise the clustering coefficient. The model produces a mean of 44 triangles and 1477 triples which is a small fraction of the 3299 triangles and 66 799 triples in the empirical network. It should be noted, however, that although the model's clustering coefficient is low relative to the empirical data, the model's clustering is still high relative to a random network with an equivalent number of interactions.

The gray band in Figure 2 shows those parameter values which produce networks with sizes within 10% of the empirical network's 5449 interactions. Those Vázquez *et al.* parameter values achieving a similar clustering coefficient and size are quite different from the empirical values. In particular, Vázquez *et al.* requires a subfunctionalization rate half of the empirical rate in order to achieve a similar size and clustering coefficient.

3.3 Heritable homomers

The results from modifying the Vázquez *et al.* (2003) model to support heritable homomers (see Section 2) are shown in Table 1. Interestingly, this model actually produces lower clustering than the original Vázquez *et al.* model. This can be attributed to the rate at which paralogous interactions are formed. Duplication events which produce paralogous interactions generate a greater increase in clustering coefficient over duplication events without paralogous interactions (Gibson and Goldberg, 2009a). While the original model adds paralogous interactions with a fixed probability after each

duplication, the heritable homomers model probabilistically adds paralogous interactions only if the duplication event involves a homomeric protein. The heritable homomers model produces a higher clustering coefficient with early duplication events, but as the network grows, the proportion of homomeric proteins in the network shrinks and the original model overtakes it in the number of paralogous interactions formed (See Supplementary Fig. 4). This is evident by observing that only 16% of the proteins in the empirical data are homomeric, while the calculated paralogous interaction rate for the original Vázquez model is 0.24. One explanation for this large discrepancy is the underreporting of homomeric interactions by large-scale assays (Gibson and Goldberg, 2009a). If a larger proportion of self-interacting proteins were identified in the empirical data, the Heritable homomers model would begin with a higher density of self-interactions which in turn would produce a higher clustering coefficient (see Seed Graph Construction in Methods of Supplementary Material).

3.4 Asymmetry

Supplemental Figure 2 shows how the clustering landscape changes when we modify the Vázquez *et al.* model to accommodate an empirically derived subfunctionalization asymmetry of 82% (see Methods of Supplementary Material). A comparison between panes A and B in the supplementary figure shows that asymmetric subfunctionalization produces higher clusterings for all parameter values.

Table 1 shows the topological properties for the Vázquez model variant run at 82% asymmetry. The topological properties indicate that introducing asymmetry improves the ability of the model to approach the properties of the empirical network (Table 1). However, the asymmetry modification still only generates half of the edges of the empirical data and less than 10% of the empirical triangles and triples.

Increasing subfunctionalization asymmetry also drives the number of components produced in the network to be more similar to that of empirical networks and is discussed in detail in the Supplementary Materials.

3.5 iSite model

Using the empirically derived network measures (Section 2.1), the iSite model generates networks with remarkable fidelity to the empirical network. Table 1 shows the mean topological values for 100 runs of the model. The iSite network size (number of edges) is within 2% of empirical. The clustering coefficient is greater than empirical: 0.162 versus 0.148, but is closer to empirical than other models. The iSite model produces roughly half of the triangles and triples observed in the empirical network. Though the triangle and triple counts are lower than the empirical values, they represent a large increase over the few triangles and triples produced by the Vázquez *et al.* model and variants.

One notable side effect of the iSite model is that it is highly asymmetric, even if subfunctionalization is symmetric (discussed in Section 4.2). If iSites are subfunctionalized symmetrically, interactions are lost with 91% asymmetry versus 82% in the empirical network. The topological properties for the modified Vázquez *et al.* model at 91% asymmetry are listed in Table 1. This increased asymmetry improves its clustering topology over other Vázquez *et al.* variants, but still lags significantly far behind both the empirical data and the iSite model.

An iSite variant which silences *all* interactions of a selected iSite with probability q produces almost complete networks and is discussed in the Supplementary Materials.

4 DISCUSSION

We have presented an alternative method of calculating model parameters which take into consideration the effect concurrent and subsequent duplications have on the model calculation. Now we discuss the advantages of our parameter estimation methodology over previous approaches. We also take a further look at the model improvements integrated into the iSite model, and how it compares to the empirical network.

4.1 Parameter estimation by network counts

The subfunctionalization rate is the proportion of redundant interactions lost between gene duplicates, and can be measured as the number of interacting neighbors not shared by both gene duplicates versus the total number of neighbors the duplicates interact with. Previously, shared (or distinct) interacting neighbors of gene duplicates have been averages of per-duplicate-pair tallies computed on empirically derived, extant, protein interaction networks (Beltrao and Serrano, 2007; Berg *et al.*, 2004; Chung *et al.*, 2006; He and Zhang, 2005; Wagner, 2001, 2002, 2003). Our whole-network approach is superior to the per-pair tally approach to finding the mean shared (or distinct) neighbors in three ways.

First, as network measures, each interaction (redundant, or lost) is given equal weight in the network regardless of the degree of the two proteins forming the interaction. This advantage parallels that of computing the clustering coefficient as a network measure (Newman, 2001). The original definition of the clustering coefficient calculated an average of each network node's neighborhood density (Watts and Strogatz, 1998). This per-node measure diminishes the clustering contribution of large degree nodes and is not well defined for nodes with degree =1.

Second, the network rate avoids complexities involving concurrent duplication. When the neighbor of a paralogous pair duplicates, there is ambiguity in assigning 'ownership' of the edge. That is, if an interaction between a paralogous pair member and one of the duplicated neighbors is lost, does the loss contribute to the subfunctionalization rate of the paralogous pair, the neighbor pair or both? Our network subfunctionalization rate calculation is indifferent to interaction 'ownership'.

Third, paralogous families born of more than one duplication event within a phylogenetic time period complicate efforts to group paralogs into pairs. In contrast to previous efforts which limited paralogy identification to pairwise relationships (Lynch and Conery, 2000), we incorporate multiple duplications intrinsic to large gene families. For each discernible time period, we measure differences in edge counts for the entire network before any duplications, after all duplications and after all subfunctionalization (Section 2.1). Supplementary Figure 6 illustrates the sensitivity of the measured subfunctionalization rate to the order in which the duplication events are selected. Our network measure avoids this sensitivity, evaluating all duplications simultaneously.

The pairwise measure also underestimates the number of potential paralogous interactions in large paralogous families. Multiple simultaneous duplications occurring in a gene family form a clique within self-interacting gene families (Supplementary Fig. 7). For d concurrent gene family duplications, pairwise enumeration of paralogous edges can identify at most d potential paralogous interactions versus the $\frac{d(d+1)}{2}$ generated via clique creation.

4.2 iSites and domains

There is a strong semantic relationship between protein domains and interaction sites (iSites). Protein domains are evolutionarily conserved portions of a protein's sequence or structure, and are associated with protein function. Similarly, iSites are evolutionarily conserved protein subunits, and the interactions that define them are analogous to protein function.

A current limitation of the iSite model is in the distribution of iSite counts in the model's network of proteins. Empirically, the majority of eukaryotic proteins have more than one domain (Apic *et al.*, 2001). The iSite model is initialized with one iSite per interaction per protein in the seed graph. Given the Erdős-Rényi random seed graph used, the distribution of iSites on each protein at model initiation is Poisson distributed ($\lambda = 3.33$). (A scale-free seed graph produces comparable results and is covered in the Supplementary Materials.) Using empirical values derived from *Saccharomyces cerevisiae* (as in Table 1), the iSite model produces proteins predominantly (89%) composed of a single iSite, despite starting with 13% single iSite proteins in the seed graph. Two- and three-iSite proteins comprise 10% of proteins and greater iSite concentrations comprise the remaining 1%. This is in stark contrast to the 22% of single-domain proteins observed in a study of *S.cerevisiae* SCOP domain distributions (Apic *et al.*, 2001). A more recent study examining both sequence- and structure-based domains achieved higher coverage of domain assignments to proteins and identified a greater proportion of single-domain proteins (35%) (Ekman *et al.*, 2005), though this is still far less than produced by the iSite model. The propensity of the model to regress to single iSite proteins is a side-effect of subfunctionalization. Supplementary Figure 8 shows how subfunctionalization reduces the number of multiple-iSite proteins.

A partial explanation for the discrepancy is that iSites are analogous to actively interacting domains, whereas domain assignments are based on similarity measures without regard to whether the domain's functionality has been partially or wholly silenced. In the model, an iSite which has lost all interactions is simply removed. Evolutionary phenomena unaccounted for in the model such as gene fusion and gene fission likely contribute to the discrepancy between the iSite model and empirical data. Compounding the difficulty in associating interactions with domains are interactions which may be detected by assays such as yeast two-hybrid but are not necessarily associated with specific domains such as phosphorylation-dependent protein interactions (Shaywitz *et al.*, 2002).

The model could also be improved by encouraging the distribution of iSites to more closely reflect that of domains observed in empirical data. The degeneration of multi-iSite proteins to single iSites as illustrated in Supplementary Figure 8 is a symptom of iSites within a protein being independent. A method encouraging multiple iSites of a single protein to remain together could mitigate the degeneration of multi-iSite proteins.

An important improvement over the Vázquez *et al.* model is in the heteromerization parameter—the probability of adding an interaction between gene duplicates. The iSite model repurposes it from probabilistic paralogous edge addition to a parameter representing selection pressure operating on heritable interactions. Notably, this closely parallels the Heritable homomers modification that was made to the original Vázquez et al. model. Despite this similarity, the iSite model achieves a high clustering coefficient unattainable by the original Vázquez model and its variants. The high clustering can be attributed to the rapid increase of interactions within iSites coupled with paralogous interactions. The iSite is the unit of redundancy subject to subfunctionalization (versus the original Vázquez model and its variants where each interaction is independently redundant). Because of this, upon duplication all interactions within each iSite will survive unchanged in either the progeny or progenitor. This concentrates interactions more quickly than the original Vázquez model. It is paralogous interactions which then provide the greatest contribution to high clustering (Gibson and Goldberg, 2009a). As paralogous interactions form among the more highly concentrated interactions within iSites, a high clustering coefficient is generated.

5 CONCLUSION

We have developed an improved methodology for estimating link dynamics rates based on empirical network data. Estimating rates through simple examination of the empirical network fails to account for the entire evolutionary history of interaction addition and loss which underlie the extant network. Our methodology attempts to 'peel back' more recent duplication events in order to more accurately assess link dynamics associated with older duplication events. Using our empirically derived rates, the Vázquez *et al.* model fails to achieve empirical clustering. A modified model integrating heritable interaction sites generates networks with clustering more closely approximating the yeast network.

ACKNOWLEDGEMENT

We would like to thank Larry Hunter and Al Goldberg for reviewing drafts of this manuscript. We would also like to thank the many constructive comments by the anonymous reviewers.

Funding: National Science Foundation under (grant MCB 0630250); National Institutes of Health training (grant T15LM009451).

Conflict of Interest: none declared.

REFERENCES

Apic,G. *et al.* (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.*, **310**, 311–325.

Beltrao,P. and Serrano,L. (2007) Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Comput. Biol.*, **3**, e25.

Berg,J. *et al.* (2004) Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol. Biol.*, **4**, 51.

Byrne,K.P. and Wolfe,K.H. (2007) Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics*, **175**, 1341–1350.

Chung,W.-Y. *et al.* (2006) Rapid and asymmetric divergence of duplicate genes in the human gene coexpression network. *BMC Bioinformatics*, **7**, 46.

Ekman,D. *et al.* (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J. Mol. Biol.*, **348**, 231–243.

Gibson,T.A. and Goldberg,D.S. (2009a) Questioning the ubiquity of neofunctionalization. *PLoS Comput. Biol*., **5**, e1000252.

Gibson,T.A. and Goldberg,D.S. (2009b) Reverse engineering the evolution of protein interaction networks. In Altman,R. B. *et al.* (eds) *Pacific Symposium on Biocomputing*, pp. 190–202.

Goldberg,D.S. and Roth,F.P. (2003) Assessing experimentally derived interactions in a small world. *Proc. Natl Acad. Sci. USA*, **100**, 4372–4376.

He,X. and Zhang,J. (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, **169**, 1157–1164.

Kellis,M. *et al.* (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, **428**, 617–624.

Lynch,M. and Conery,J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.

Newman,M.E. (2001) The structure of scientific collaboration networks. *Proc. Natl Acad. Sci. USA*, **98**, 404–409.

Newman,M. *et al.* (eds) (2006) *The Structure and Dynamics of Networks: (Princeton Studies in Complexity)*. Princeton University Press, Princeton, New Jersey.

Pržulj,N. *et al.* (2004) Modeling interactome: scale-free or geometric? *Bioinformatics*, **20**, 3508–3515.

Shaywitz,A.J. *et al.* (2002) Analysis of phosphorylation-dependent protein-protein interactions using a bacterial two-hybrid system. *Sci. STKE*, **2002**, pl11.

Solé,R.V. *et al.* (2002) A model of large-scale proteome evolution. *Adv. Comp. Syst.*, **5**, 43.

Vázquez,A. *et al.* (2003) Modeling of protein interaction networks. *ComPlexUs*, **1**, 38–44.

Wagner,A. (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol*., **18**, 1283–1292.

Wagner,A. (2002) Asymmetric functional divergence of duplicate genes in yeast. *Mol. Biol. Evol.*, **19**, 1760–1768.

Wagner,A. (2003) How the global structure of protein interaction networks evolves. *Proc. Biol. Sci.*, **270**, 457–466.

Watts,D.J. and Strogatz,S.H. (1998) Collective dynamics of ‘small-world’ networks. *Nature*, **393**, 440–442.