Pacific Symposium on Biocomputing 15:178-189(2010)

GEOMETRIC EVOLUTIONARY DYNAMICS OF PROTEIN INTERACTION NETWORKS

NATAŠA PRŽULJ*, OLEKSII KUCHAIEV, ALEKSANDAR STEVANOVIĆ, and WAYNE HAYES

Department of Computer Science,
University of California, Irvine
CA, 92697-3425, USA
E-mail: natasha@ics.uci.edu

Understanding the evolution and structure of protein-protein interaction (PPI) networks is a central problem of systems biology. Since most processes in the cell are carried out by groups of proteins acting together, a theoretical model of how PPI networks develop based on duplications and mutations is an essential ingredient for understanding the complex wiring of the cell. Many different network models have been proposed, from those that follow power-law degree distributions and those that model complementarity of protein binding domains, to those that have geometric properties. Here, we introduce a new model for PPI network (and thus gene) evolution that produces well-fitting network models for currently available PPI networks. The model integrates geometric network properties with evolutionary dynamics of PPI network evolution.

Keywords: evolutionary dynamics, protein-protein interaction networks, geometric network model

1. Introduction

Understanding protein-protein interactions (PPIs) and the complex networks that they form is a central problem in systems biology. The crucial role that proteins play in most cellular processes justifies intense study of the complex wiring of their interactions. Note that the network of interactions may contain significant information that cannot be extracted solely from a study of protein sequences, because individual proteins rarely function alone. Instead they "cooperate" with other proteins, and the resulting complex networks of protein-protein interactions may provide information that cannot be ascertained from the study of sequences or even sequence comparison. Hence, in the post genomic era, there is considerable scientific value in understanding topological properties of PPI networks and the biological origin and meaning of those properties.

The mathematical foundations for studying PPI networks are graph theory and statistics. A PPI network is modeled as an undirected unweighted *graph* (also called a *network*) $G(V, E)$, where V is a set of nodes (proteins) and E is a set of edges (i.e., interactions between protein pairs). Since self-loops in E significantly complicate the analysis of graphs, we make the simplifying assumption that they are not allowed. Since subgraph isomorphism is NP-complete,[1] exact network comparisons are computationally infeasible. Thus, easily computable heuristics are used for comparing networks. These heuristics are commonly called *network properties* and can historically be divided into two groups: *local* and *global* properties. Global properties include the *degree distribution*, the *clustering coefficient*, the *average shortest path length*, and various forms of *network centralities*.[2,3] Local ones include *network motifs*[4–6] and *graphlets*,[7] both of which relate to the occurrence of small subgraphs in a larger graph. While motifs refer to subgraphs (not necessarily induced) that occur at an unusually high frequency, graphlets refer to *all* induced subgraphs regardless of their frequency. As such, graphlets provide a more thorough description of a network and allow their comparison, such as through *graphlet degree distribution agreement (GDD-agreement)*;[8,9] see section 2.2 for details. By using these properties to compare networks, well-fitting network models for biological networks have been proposed.[7,8,10,11] Also, network properties have been used to suggest protein function and involvement in disease.[12–16] Network models have further been exploited to guide biological experiments and discover new biological features.[17,18]

[*] Corresponding author.

The first attempts to model real-world networks began with the Erdös-Rényi (ER) random graph model.[19] This model has well-studied mathematical properties, although it is very simplistic. Erdös-Rényi random graphs have only two parameters: the number of nodes in the network and the probability p of an edge between any two nodes. The ER model poorly captures both global and local properties of PPI networks.[7,8] This implies that the structure of PPI networks is not completely random, presumably because evolution has imposed structural patterns that deserve close investigation.

Scale-free networks, popularized by Barabási and Albert, were proposed as a new and better model for early data sets of protein-protein interaction networks.[20] In these networks, degree distributions follow a power-law. The popularity of this model is explained by the fact that in many real-world networks (including PPI networks) a part of the degree distribution obeys a power-law. Barabási and Albert also proposed a *preferential attachment* model, which can generate networks with power-law degree distributions. The scale-free model, however, has several conceptual drawbacks. First, the degree distribution is not a sufficiently discriminative measure, since two networks can have exactly the same degree distribution, but completely different local structure. For example, a graph containing 100 triangles has exactly the same degree distribution as one containing a single 300-node cycle. In fact, the *gene duplication and mutation model* for generating networks with power-law degree distributions, proposed by Vaźquez et al.,[21] fits PPI networks much better than does the preferential attachment model, even though both generate graphs with power-law degree distributions. Furthermore, the systematic study of the fly's PPI network classified it as a duplication-mutation-complementation network instead of a preferential attachment one.[22] Thus, variants of the scale-free model of gene duplications and mutations have been proposed.[23–25] The second major drawback of the scale-free network model for PPI networks comes from the fact that currently available PPI datasets are both noisy and incomplete. It has been shown that subsets of scale-free networks are not scale-free[26,27] and therefore, since current PPI datasets have power-law degree distributions, it is not clear that complete and clean PPI networks are scale-free. Recently, *geometric random graphs*[28] were introduced as a better model of PPI networks; in these graphs, nodes correspond to points in space distributed uniformly and independently at random and edges exist between nodes corresponding to points that are close in space; see section 2.1 for details. These graphs have Poisson degree distributions, but there exist variants of geometric graphs in which this is not the case. In has been shown that geometric random graphs fit PPI networks much better than other commonly used network models despite the fact that parts of degree distributions of PPI networks often follow a power-law.[7,8,11,29]

In this paper we introduce a new model of PPI network evolution that results in networks that provide the best currently known fit to high confidence PPI networks. Since geometric *random* graphs seem to provide the best fit to the currently available PPI networks and since genomes have evolved through gene duplication and mutation events rather than at random, we bridge the concepts of network geometricity with the evolutionary dynamics. We introduce two new network models of *geometric gene duplication and mutation*, which utilize geometric graph principles to model the evolutionary dynamics of PPI networks.

2. Methods

It is important to distinguish between two conceptually different types of network models, which we refer to as *descriptive* and *network-driven* models. Descriptive models describe general properties of all networks of a particular type (e.g., PPI networks). For example, the scale-free model reproduces the power-law degree distribution (regardless of the exponent γ in the power law $P(k) \sim k^{-\gamma}$) which was observed in many, though not all, PPI datasets. Our new geometric gene duplication models and scale-free gene duplication[21] models are also descriptive because they model the principle (gene duplication and mutation) by which all PPI networks have evolved. A network-driven model, in contrast, tries to model a *particular* PPI network instance as well as possible. For example, trained geometric model ("geo-train"),[11] stickiness-index-based model ("sticky")[10] and Erdös-Rényi random graphs with the same degree distribution as data ("er_dd") require a *particular* network example and then try to reproduce its structure.

Since descriptive models do not need to be fitted to the particular network example they do not suffer from over-fitting. On the other hand, network-driven models must have an example network to learn their parameters and therefore might over-fit the data. Hence, to avoid over-fitting, one should be careful and properly adjust the number of free parameters in such models using, for example, the Bayesian Information Criterion.[11] In this study, we show how the geometric graph framework can be used to create a well-fitting descriptive model of PPI networks. For a geometric network-driven model, see Kuchaiev and Pržulj (2009).[11]

2.1. *Geometric Gene Duplication and Mutation Models*

A *geometric random graph*[28] is a graph $G(V, E)$ with the set of nodes V distributed uniformly at random in some metric space. An edge exists between two nodes u and v if the distance between them $d(u, v)$ is less than ϵ, for some constant ϵ and some appropriate norm. The crucial parameters of this model are: the metric space, its dimensionality, and the distribution of nodes in that space. Intuitively, each protein can be described with its biochemical properties and therefore proteins reside in some multidimensional biochemical space. However, currently, it is hard even to hypothesize about the nature or dimensionality of that space. In this study, we focus on altering the distribution of the nodes in a low-dimensional Euclidean space in a way which simplistically models evolutionary dynamics of protein-protein interaction networks. Euclidean space is chosen just as a proof of concept. Note that highly dimensional spaces are less interesting, since if we allow enough dimensions, we can trivially embed a network into such a space. Thus it is encouraging to discover that only a few dimensions are required to accurately model this space geometrically,[7,8,11,29] indicating that geometricity is an important factor in modeling PPI networks. Furthermore, PPI networks can be directly embedded into a low dimensional space, in the sense that nodes sharing a graph-theoretic neighborhood can be placed close together in a geometric space in such a way that the resulting geometric graph is almost identical to the original.[29]

We introduce two geometric network models that incorporate the principles of gene duplications and mutations. Each of our models determines the principle by which the network is grown from an initial, small seed network. Growth is governed by adding new nodes intended to model gene duplications and mutations, moderated by natural selection as follows. A duplicated gene starts at the same point in biochemical space as its parent, and then "evolutionary optimization" acts either to eliminate one, or cause them to slowly separate in the biochemical space. This means that the child inherits some of the neighbors of its parent while possibly gaining novel connections as well. The further the "child" is moved away from its "parent," the more different their biochemical properties are. The randomness in the direction of the move models which subset of parent's properties (i.e. network interactions) will be preserved.

We refer to our two new models as *GEO-GD expansion* and *GEO-GD with a probability cutoff.* Each GEO-GD model network starts from a small initial *seed* network. For simplicity we use a 5-node clique, although we do not know what effect the exact makeup of the seed network will have on the final structure (this will be tested in a future paper). To create our seed networks, we place 5 nodes uniformly at random inside a sphere of radius $\frac{\epsilon}{2}$ in the embedding space. As a proof of concept, we use 3-dimensional Euclidean space for growing both of our GEO-GD models.

2.1.1. *GEO-GD Expansion Model*

Starting from the seed network, this model adds nodes iteratively, by choosing as the parent an existing node uniformly at random and placing a child node in a random direction at a randomly chosen distance of at most 2ϵ from the parent, where ϵ is the same parameter as was used in the definition of a geometric random graph. The movement at a distance less than ϵ allows the child to keep some of the parent's connections, whereas the movement at a distance of greater than ϵ allows the child to form a completely new set of connections. Thus, the child-node is adjacent to some of the neighbors of the parent-node, while at the same time potentially gains new interactions. We stop once we reach a predetermined number of nodes.

2.1.2. *GEO-GD with a Probability Cutoff Model*

This model is almost identical to the expansion model, except that the child can be duplicated in two different ways, rather than just one. In both cases, the child moves in a randomly chosen direction, but the two cases differ in the distance the child can move: with probability p, the child can move a maximum distance of ϵ, while with probability $1-p$ it can move up to 10ϵ; the second case is meant simply to model a large mutation rather than a small one.

Figure 1 presents some examples of distributions of points on the plane generated by our models. As it follows from this figure, the GEO-GD with a probability cutoff model generates networks with more pronounced clusters than GEO-GD expansion model, especially for higher values of parameter p. We only present points in the figure without edges, since edges would clutter the figure. Note that after n nodes were generated by *GEO-GD Expansion* or *GEO-GD with a Probability Cutoff* model, the resulting number of edges E' in the model network might be different from the number of edges in the data. Hence, in order to tune this number to our desired number of edges (i.e., to the number of edges of the PPI network that we are modeling) we have to rescale our initial ϵ by a small amount. If we denote by E the number of edges in the data, we first calculate the distances between all pairs of nodes in the model network, then we order them from the smallest to the largest and choose the E^{th} smallest distance as the re-scaled value $\epsilon^{'}$. Then we connect two nodes by an edge if they are closer than $\epsilon^{'}$ in the space. Since we rescale the ϵ in order to produce as many edges as in the data, the initial value of ϵ is not important.

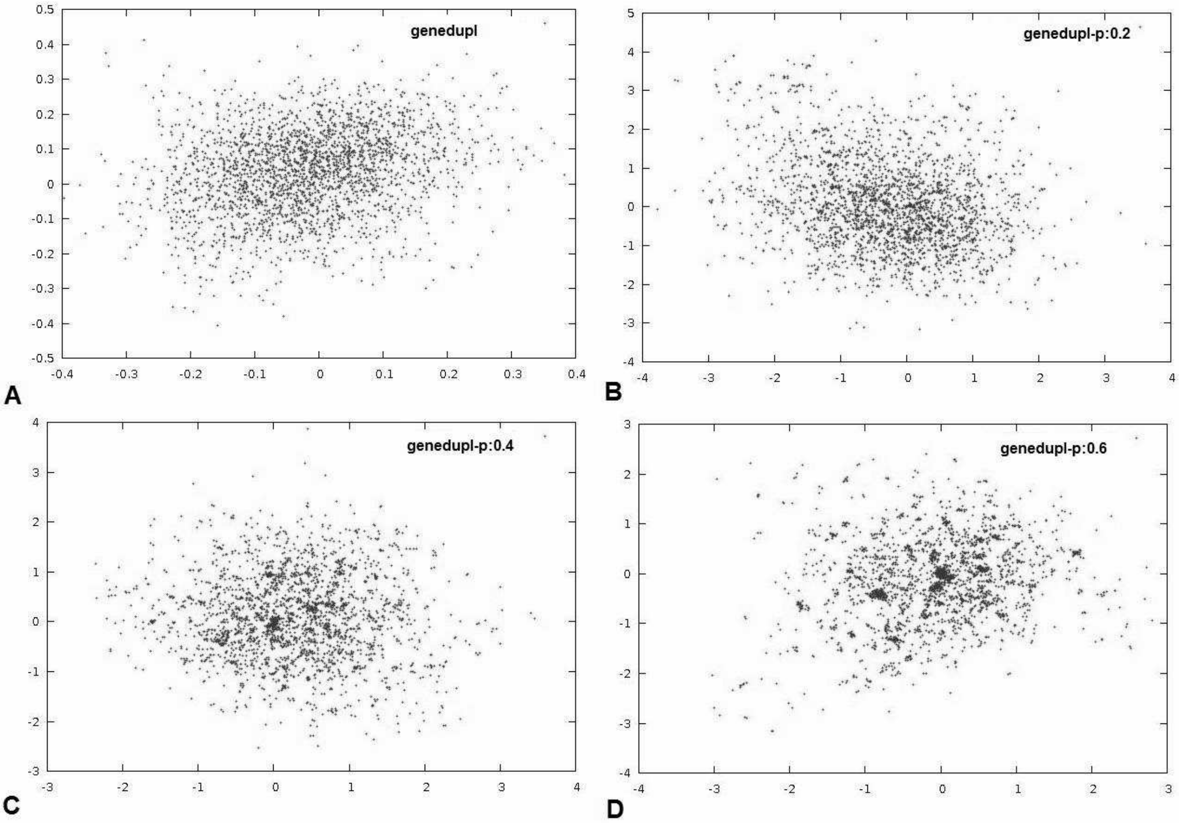

Fig. 1. Example distributions of points in the 2 dimensional Euclidean space, generated by (A) GEO-GD expansion model (genedupl), and GEO-GD with a probability cutoff model where (B) (genedupl-p:0.2) $p = 0.2$, (C) (genedupl-p:0.4) $p = 0.4$, and (D) (genedupl-p:0.6) $p = 0.6$ (genedupl-p:0.6). Here p is the probability of the move within ϵ distance from the parent node and $1-p$ is the probability of the move within 10ϵ distance.

2.2. *Evaluation of the Models*

To evaluate the fit of the network model to the data, we need to compare the model networks to the PPI networks. As described in the Introduction, since large network comparisons are computationally infeasible, we examine the fit of local and global network properties of the model to the data. Since PPI networks are incompletely explored, global properties of such incomplete data are likely to be biased, or even misleading with respect to the currently unknown complete PPI networks. However, certain parts of these networks are very well studied (e.g., parts relevant for human disease). Thus, since we have detailed knowledge of certain local parts of PPI networks, but the data outside these well-studied parts are currently incomplete, global statistics are likely to provide misleading information about the PPI network as a whole, whereas local statistics are likely to be valid and meaningful. For these reasons, we focus on examining the similarity of networks by using our highly constraining measure of local network similarity, *graphlet degree distribution (GDD) agreement*.[8] Its formal definition is as follows.

First, a *graphlet* is a small, connected, induced subgraph of a network;[7] an *induced* subgraph with a node set $A \subseteq V$ of a graph $G(V, E)$ is obtained by taking A and all edges of G having both endpoints in A. There are 30 possible non-isomorphic graphlets on 2, 3, 4 and 5 nodes[7] (Figure 2; isomorphism is defined below). The GDD-agreement is a generalization of the degree distribution. Since the degree distribution $P(k)$ measures the number of nodes "touching" k edges and since an edge is the only 2-node graphlet, we generalize the degree distribution into the spectrum of distributions measuring the number of nodes "touching" k graphlets, for each of the 30 2-5-node graphlets. Clearly, the degree distribution is the first one in this spectrum.

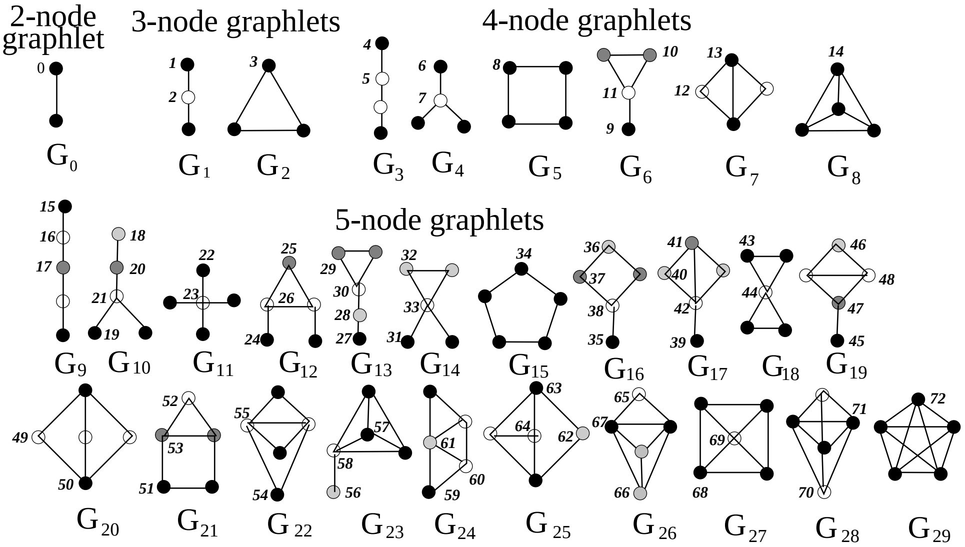

Fig. 2. The 30 2-5-node graphlets $G_0, G_1, G_2, \ldots, G_{29}$. In each graphlet, nodes belonging to a different automorphism orbit are of different shade. The 73 automorphism orbits of the 30 graphlets are labeled from 0 to 72. The figure is taken from Pržulj (2007).[8]

However, when we do this, we notice that while an edge was "symmetric," other graphlets might not be in the following sense. From a topological point of view, it is relevant to distinguish between *automorphism orbits* within each graphlet. For example, in a 3-node linear path (graphlet G_1 in Figure 2), the "end-nodes" are topologically identical, i.e., can be mapped to each other by an *automorphism*, an isomorphism of a graph with itself, where an *isomorphism* between two graphs is a bijection of their node sets that preserves their adjacency; the "middle node" of a 3-node path can only be mapped to itself by an automorphism. Therefore, a 3-node path has two different automorphism orbits. There are 73 automorphism orbits for the 30 2-5-node graphlets (Figure 2).[8] Thus, the *graphlet degree distribution (GDD)* is a 73-component distribution of a network. Its j^{th} component, $d^j(k)$, is the sample distribution of the number of nodes in the network touching a particular graphlet k times at automorphism orbit j. The *GDD-agreement* is a similarity measure between graphlet degree distributions of two networks. It is a number between 0 and 1, meaning that two networks have similar GDDs if their GDD-agreement is close to 1, and otherwise, their GDDs are different (see Pržulj (2007)[8] for details). Note that GDD-agreement is a very strong measure of structural similarity of small-world networks (those with small diameters), since in these networks 5-node graphlets reach most of the network from each node. Since PPI networks are small-world, as demonstrated in Table 1 below, GDD-agreement based on up to 5-node graphlets is a strong measure of comparing their topologies to each other, or to model networks. We use our GraphCrunch software package[9] to calculate GDD-agreement and other network properties and evaluate the fit of model networks to the data.

3. Results

We test our model on four eukaryotic organisms whose PPI network data were produced using different biotechnologies, as well as deposited into curated databases. The organisms are baker's yeast *Saccharomyces cerevisiae*,[33–35] fruitfly *Drosophila Melanogaster*,[36,37] worm *Caenorhabditis elegans*[38,39] and human.[38,40–42] Table 1 shows the PPI networks that we analyze along with their sizes, some global network properties, and references the data originates from. YH1 and YH2 contain high-confidence parts of Collins *et al.*[33] and von Mering *et al.*[34] data sets that contain both binary interaction and co-complex data (see section 4.3 for details). Similarly, FH1 and FH2 contain high confidence parts of Giot *et al.*[36] and Finley *et al.*[37] fruitfly binary PPI data sets; FH2 is more recent and thus believed to be of higher quality than FH1. WE1 is the worm PPI network downloaded from BioGRID and WH1 is the high confidence binary PPI network from Simonis *et al.*[39] Human PPI network HE1 contains binary interactions from Rual *et al.*,[40] HH1 contains high-quality binary interactions from Venkatesan *et al.*,[42] while HE2 and HE3 are downloaded from BioGRID and HPRD, respectively. All of the networks have short average pathlenghts, i.e., they are small-world networks.

Table 1. PPI Networks that we analyze.

Network	Organism	Number of Nodes	Number of Edges	Avg Path-length	Clustering Coef.	Reference (source of the data)
YH1	Yeast	1,622	9,074	5.53	0.55	Collins *et al.*[33]
YH2	Yeast	988	2,455	5.19	0.34	von Mering *et al.*[34]
YH3	Yeast	2,018	2,705	5.61	0.04	Yu *et al.*[35]
FH1	Fly	4,602	4,637	9.43	0.01	Giot *et al.*[36]
FH2	Fly	1,345	3,112	4.50	0.03	Finley *et al.*[37]
WE1	Worm	2,821	4,470	4.84	0.02	BioGRID (v2.0.51)[38]
WH1	Worm	2,528	3,706	5.32	0.02	Simonis *et al.*[39]
HH1	Human	235	239	4.53	0.00	Venkatesan *et al.*[42]
HE1	Human	1,873	3,463	4.34	0.03	Rual *et al.*[40]
HE2	Human	8,446	25,525	4.63	0.10	BioGRID (v2.0.51)[38]
HE3	Human	9,182	34,119	4.26	0.10	HPRD(v7)[41]

We compare the fit of our new models to PPI networks with the fit of other commonly used network models. The list of network models that we analyze is presented in Table 2. For each model, we evaluate the fit to the data to 30 random networks from the model that are of the same size as the data (have the same number of nodes and edges as the data). We report the averages and standard deviations of their fit in Figures 3, 4, 5, and 6 below. For our GEO-GD model with cutoff probability p, we vary p over all possible values between 0.1 and 0.9 in increments of 0.1 to determine p that yields the best fit. For our trained geometric model,[11] for each of the species, we trained it on the part of the species' PPI network that is reported to be of high confidence. The scale-free gene duplication model that we analyze[21] (denoted by "vespdd:x," see Table 2) in addition to the number of nodes and edges has two probabilistic parameters p and q representing the probabilities of a child node to keep the parent's interactors and form new ones, respectively. Similar to what we do for our GEO-GD model to generate the best fitting model networks to the data, we vary p for vespdd:x from 0.3 to 0.7 in increments of 0.1 and for each p seek q using a binary search starting from $q = 0.5$ such that the resulting number of edges in the model network is within 1% of the number of edges in the data network (the number of nodes is the same as in the data). Figures 3, 4, 5 and 6 present GDD-agreements between the data and the model networks.

Table 2. Network models that we analyze.

Model Abbreviation	Network Model
er	Erdös-Rényi (ER) random graph model[19]
er_dd	ER model with the same degree distribution as the data[43]
geo	Geometric random graph model[7]
sf	Scale-free Barabási-Albert preferential attachment model[20]
sticky	Stickiness-index-based model[10]
geo-trained	Trained geometric model[11]
vespdd:x	Scale-free gene duplication model[21] with probability $p = x$ of child node keeping parent's interactors
genedupl	GEO-GD expansion model
genedupl-p:x	GEO-GD with a probability cutoff model, with cutoff probability $p = x$

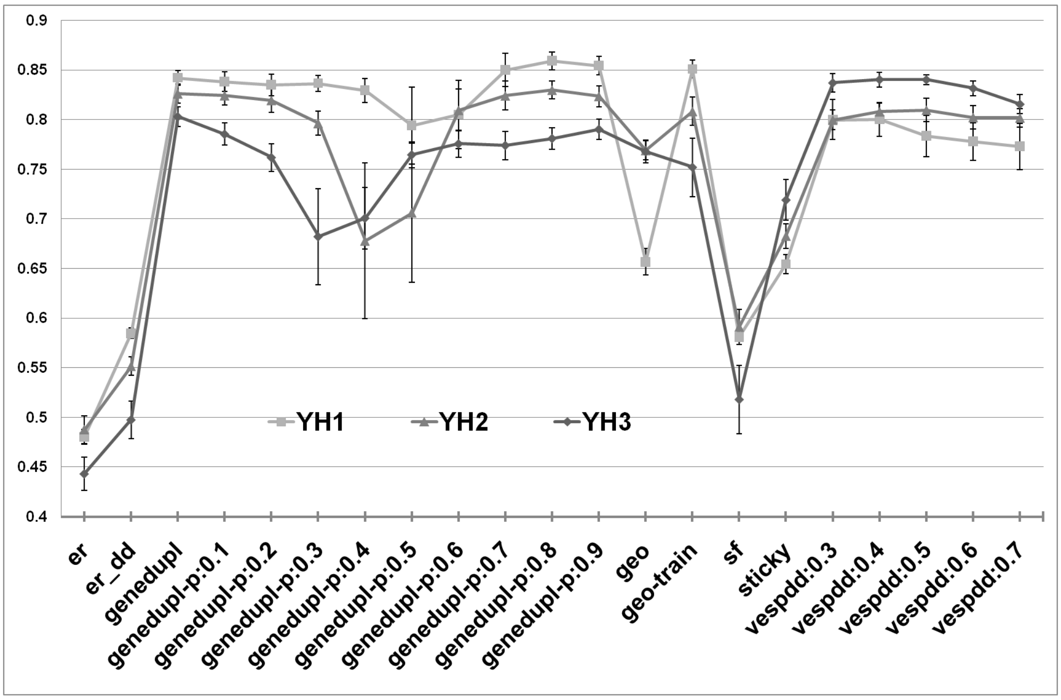

Fig. 3. GDD-agreement between yeast PPI and model networks. $x-$axis presents different model networks (see Table 2), $y-$axis presents the average values of GDD-agreements between the data and 30 model networks from each model. Lines with different point shapes correspond to different PPI networks (see Table 1); the error bar around a point is one standard deviation above and below the mean.

For yeast YH1 and YH2 PPI networks, our new GEO-GD model outperforms any other model, with the best results being achieved by the GEO-GD model with the probability cutoff $p = 0.8$. For YH3 yeast network, vespdd:x model slightly outperforms any of the geometric models. Note that YH3 network is much sparser than YH1 and YH2 in terms of the number of edges and it has a much smaller clustering coefficient than the other two yeast networks (Table 1). Thus, it is possible that YH3 contains many false negatives, i.e., missing interactions and that vespdd:x model is better for modeling sparser networks. We comment on the types of interactions (binary versus co-complex) of these three networks in the Discussion section. Trained geometric model also provides a good fit to the data, but performs slightly worse than the best geometric and scale-free gene duplication models (Figure 3). Since yeast PPI networks are currently the most complete over all eukaryotic PPI networks, the superior fit of our GEO-GD model suggests that our model successfully captures the evolutionary dynamics of PPI networks.

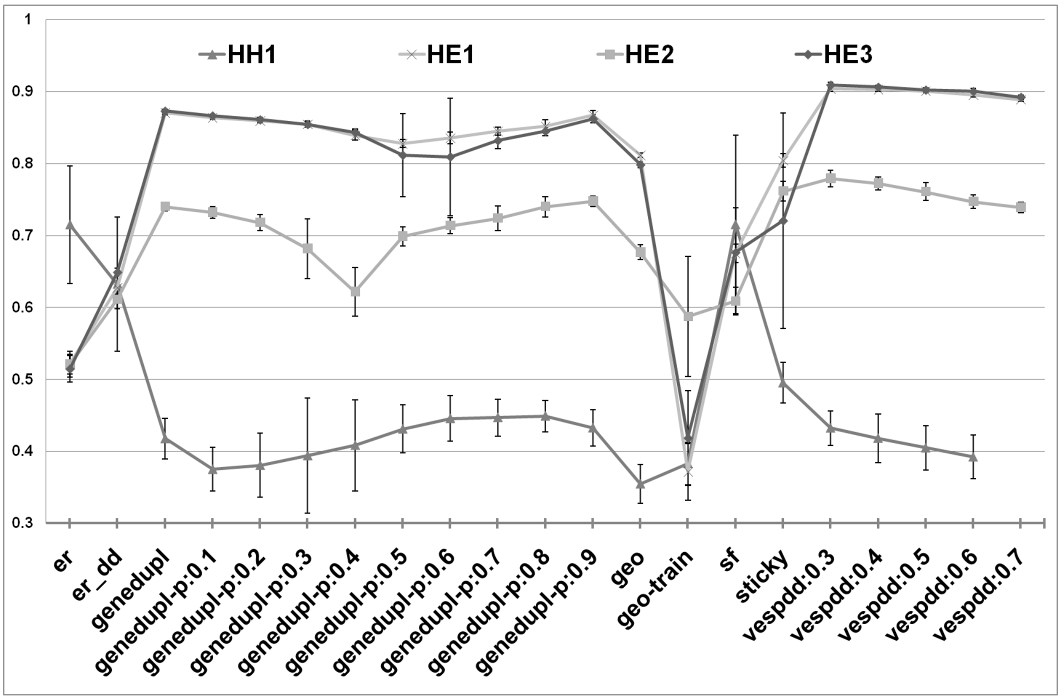

Fig. 4. GDD-agreement between human PPI and model networks. See the legend of Figure 3.

The situation is slightly different for PPI networks of human, fruitfly, and worm where scale-free gene duplication model slightly overperforms the geometric ones (Figures 4, 5 and 6). For human PPI networks, in all cases the best model is scale-free gene duplication model, except for HH1 network for which the best model is scale-free preferential attachment model (sf). Note however, that HH1 human network that is very small and sparse containing only 239 interactions between 235 proteins. Also, as mentioned above, all PPI networks of human are less complete than the PPI networks of yeast. For example, even though HH1 network contains only high-quality binary human PPIs (obtained by yeast-two-hybrid, Y2H, experiments), it is clearly an extremely small sample from the full human interactome. Also, since HH1 is extremely sparse, this indicates that many true interactions are likely to be missing. This is further corroborated by the fact that Erdös-Rényi random graphs provide the best fit to HH1 (Figure 4) and it is widely believed that topology of PPI networks is not random. Furthermore, the scale-free preferential attachment model (sf) provides as good of a fit to HH1 as Erdös-Rényi random graphs. This supports the hypothesis of the scale-free nature of incompleteness in the data.[8,26,27] Geometric gene duplication models (expansion and probability cutoff) have improved their fit over the geometric random graph model, only slightly doing worse than scale-free gene duplication model (we comment on the meaning of this slight underperformance in the Discussion section).

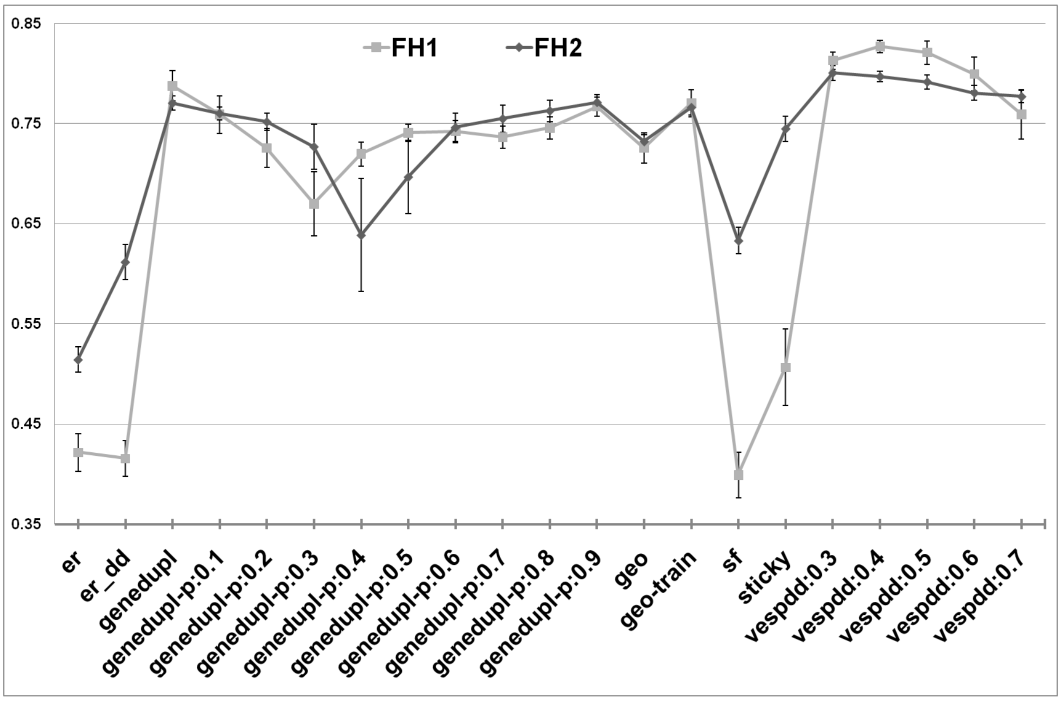

Fig. 5. GDD-agreement between fruitfly PPI and model networks. See the legend of Figure 3.

For fruitfly and worm PPI networks, scale-free gene duplication model slightly overperforms our geometric gene duplication models, while our gene duplication models outperform all other models (Figures 5 and 6).

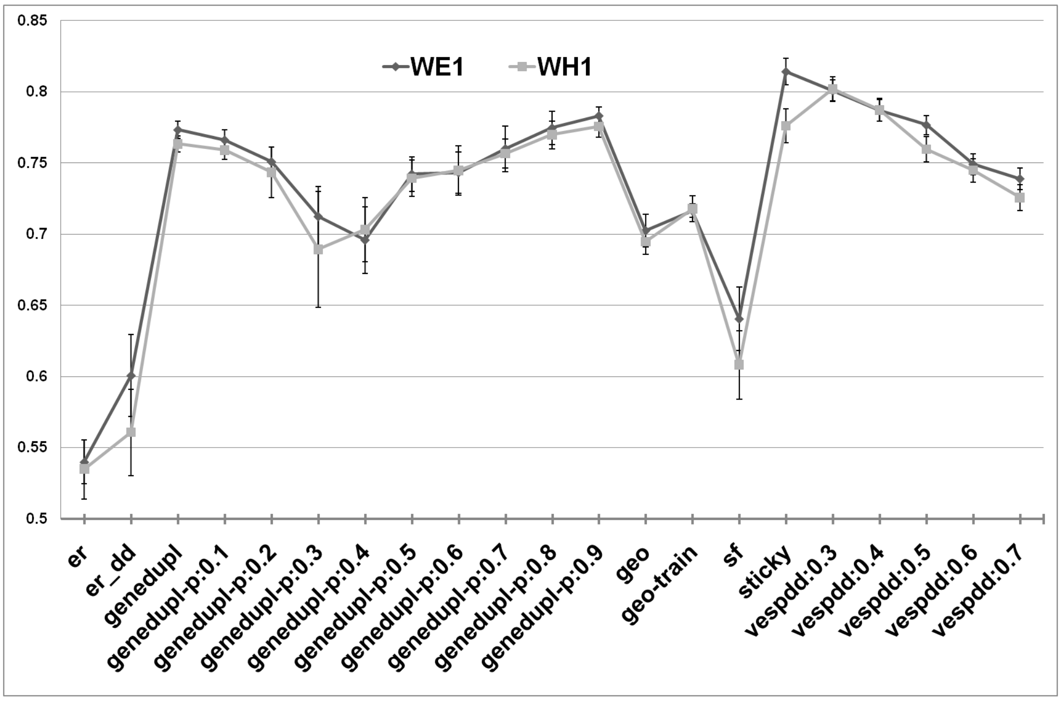

Fig. 6. GDD-agreement between worm PPI and model networks. See the legend of Figure 3.

4. Discussion

4.1. *Sensitivity of GDD-agreement*

Note that model networks generated with the same parameters that are coming from any random graph model that we analyzed have GDD-agreement of 0.86 ± 0.01.[8] Thus, since GDD-agreement of YH3 with our GEO-GD genedupl model (Table 2) is 0.8 and its GDD-agreement with the vespdd:x model is 0.84, the difference in the fit of these two models to the data falls within the sensitivity error of the GDD-agreement measure. Similar holds for the slight overperformace of our GEO-GD over vespdd:x for YH1 and YH2 networks. That is, both GEO-GDD and vespdd:x model provide the best possible fit to the yeast PPI networks that can be measured with the best currently available network comparison tool. Our contribution is not only in proposing an intuitive, geometric paradigm of evolutionary network dynamics, but also in designing such models with fewer parameters than similar scale-free-based models (see below).

4.2. *Parameters of Network Models*

An important characteristic of any model is its number of parameters. For any network model, we need at least two parameters (specified either explicitly or implicitly): the number of nodes and the number of edges in the network. More complicated models can have additional parameters. Clearly, among two models which fit the data equally well, the model with fewer parameters is better, since it provides a simpler description of the phenomena. The scale-free gene duplication model,[21] in addition to the number of nodes and edges, has two parameters p and q which are the probabilities of keeping old and forming new connections, respectively, for the duplicated node. In contrast, our new geometric gene duplication with a probability cutoff model has only one additional parameter, p, whereas geometric gene duplication expansion model does not have any parameters except of the number of nodes and edges in the network. As our experiments show, our new models fit denser high quality yeast PPI networks better than any other network model. Also, they perform on the networks of other species approximately the same as scale-free gene duplication model. The slightly better performance of a scale-free gene duplication model on human, fly and worm networks can be explained by the lower quality of these networks (compared to yeast networks) and the fact that this model has more parameters than our new models. Since our new geometric gene duplication model has fewer parameters, we can conclude that it is a better network model. Of course, one might argue that geometric graphs have an additional parameter, D, the dimensionality of the underlying Euclidean space. However, our previous research shows that the exact value of dimensionality is not important.[7,8,11,29] Instead, the important fact is that the PPI networks are well modeled by *low-dimensional* geometric graphs.

4.3. *Types of PPI Data*

We need to distinguish between PPI networks containing solely binary interactions (obtained by Y2H) and those containing co-complex data (obtained by mass spectrometry of purified complexes). Since the “spoke” and “matrix” models are used for co-complex PPIs, binary interaction networks are believed to have fewer false positives than networks containing co-complex data.[39,42] In the “spoke” model, edges exist between the bait and each of the preys, but not between the preys, while in the “matrix” model, a fully connected graph is formed between the bait and all preys. However, binary data still contain false negatives (missing interactions) due to technological limitations of Y2H (e.g., Y2H is not good at detecting membrane PPIs). The goal of our study is to provide a better model for high confidence PPI networks that are as complete as possible. Our new GEO-GD model outperforms any other model for yeast YH1 and YH2 networks, which are high confidence parts of the yeast interactome containing both binary and co-complex data, although vespdd:x model is a close contestant (Figure 3). YH3 network contains only binary interactions detected by high quality Y2H experiments. However, this network is extremely sparse, containing only 2,705 interactions amongst 2,018 proteins and therefore, it contains many false negatives. Nevertheless, the fit of our GEO-GD to YH3 network is remarkable. It is possible that if we used as the seed network a sparse graph, e.g. a 5-

node path, instead of a 5-node clique for growing our GEO-GD models, we would obtain even better fitting GEO-GD networks for binary interaction data.

Human PPI network HE1 consists solely of binary interactions. HH1 network consists only of high-quality binary interactions, but its level of false negatives (missing interactions) is very high, which potentially explains why simple Erdös-Rényi random graphs (er) and scale-free preferential attachment models (sf) are the best for this data set (Figure 4). HE2 and HE3 human networks were downloaded from BioGRID and HPRD and thus contain both types of PPIs, which makes them more complete, but with higher levels of false positives than in HE1 and HH1. Both of the fly PPI networks contain only binary PPIs and they can both be modeled indistinguishably well by geometric and scale-free gene duplication and mutation models (Figure 5). Worm WH1 network contains solely binary interactions, whereas WE1 contains both types of PPIs. They are both modeled well by geometric as well as by scale-free gene duplication and mutation models (Figure 6).

5. Conclusion

We have shown how the geometric graph framework can be used to model the principle of gene duplications and mutations by which all PPI networks have evolved. We demonstrated that our new descriptive network models of geometric evolutionary dynamics are well-fitting to the currently available PPI networks of eukaryotic organisms. The fact that geometric and scale-free gene duplication and mutation models always perform approximately the same and often outperform other models leads to the conclusion that it is not the power-law degree distribution of the currently available PPI datasets that is important, but the underlying processes of evolutionary dynamics that created these networks.

A mathematical model of any real-world phenomena has two ultimate goals: to provide better understanding of the phenomena and to allow practical applications. The scale-free models were used to describe the data; however their practical applications were limited to simply estimating the size of the interactomes. Geometric framework allows to work with network's nodes as with point in the metric space which is much more convenient from the mathematical point of view and can be used for such important practical applications as, for example, PPI network de-noising.[18] Due to their better fit to the data and mathematical convenience for the practical applications, we believe that geometric random graph model is the most promising framework for working with PPI networks.

Acknowledgments

This project was supported by the NSF CAREER grant IIS-0644424.

References

1. S. A. Cook, *Proc. 3rd Ann. ACM Symp. on Theory of Computing*, 151-158 (1971).
2. M. E. J. Newman, *SIAM Review*, **45**, 167-256 (2003).
3. M. E. J. Newman, in *The New Palgrave Encyclopedia of Economics*, L. E. Blume and S. N. Darlauf (eds.), **2**nd edition (2008).
4. R. Milo, S. S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, *Science*, **298**, 824-827 (2002).
5. S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, *Nature Genetics*, **31**, 64-68 (2002).
6. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, *Science*, **303**, 1538-1542 (2004).
7. N. Pržulj, D. G. Corneil, and I. Jurisica, *Bioinformatics* **20**, 3508-3515 (2004).
8. N. Pržulj, *Proceedings of the 2006 European Conference on Computational Biology (ECCB'06), Bioinformatics*, **23**, e177-e183 (2007).
9. T. Milenković, J. Lai, and N. Pržulj, *BMC Bioinformatics*, **9**, 70 (2008).
10. N. Pržulj and D. J. Higham, *Journal of the Royal Society Interface*, **3**, 10 (2006).
11. O. Kuchaiev and N. Pržulj, *Proceedings of the 2009 Pacific Symposium on Biocomputing*, 39-50 (2009).
12. R. Sharan, I. Ulitsky, I. and R. Shamir, *Molecular Systems Biology*, **3:88** (2007).
13. T. Milenković and N. Pržulj, *Cancer Informatics*, **2008:6**, 257-273 (2008).

14. C. Guerrero, T. and Milenković, N. and Pržulj, J. J. Jones, P. Kaiser, and L. Huang, *PNAS*, **105**, 13333-13338 (2008).
15. T. Milenković, V. Memišević, A. K. Ganesan, and N. Pržulj, *Journal of the Royal Society Interface*, to appear (2009).
16. R. Aragues, C. Sander, and B. Oliva, *BMC Bioinformatics*, **9:172** (2008).
17. M. Lappe and L. Holm, *Nature Biotechnology*, **22**, 98-103 (2004).
18. O. Kuchaiev, M. Rašajski, D. J. Higham, N. Pržulj, *PLoS Computational Biology*, **5(8)**, e1000454 (2009).
19. P. Erdös, and A. Rényi, *Publ. Math.* **6**, 290-297 (1956).
20. A.L. Barabasi and R. Albert, *Science* **286**, 509-512 (1999).
21. A. Vaźquez, A. Flamminia, A Maritana, A. Vespignani, *Complexus*, **1**, 38-44 (2003).
22. Manuel Middendorf, Etay Ziv, Chris H. Wiggins , *PNAS* **102:9**, 3192-3197 (2005).
23. R. V. Sole, R. Pastor-Satorras, E. Smith, T. Kepler, *Adv. Complex Syst.*, **5**, 43-54 (2002).
24. A. Wagner, *Proc. R. Soc. London B* **270**, 457-466 (2003).
25. F. Chung, L. Lu, T. Dewey, D. Galas, *J. Comput. Biol. 10*, **5**, 677-688 (2003).
26. M.P.H. Stumpf, C. Wiuf, and R.M. May, *PNAS* **102**, 4221-4224 (2005).
27. J. D. H. Han, D. Dupuy, M. Bertin, M. E. Cusick, M. and Vidal, *Nature Biotechnology*, **23**, 839-844 (2005).
28. M. Penrose. Geometric Random Graphs, *Oxford University Press* (2003).
29. D. J. Higham, M. Rašajski, and N. Pržulj, *Bioinformatics*, **24**, 8 (2008).
30. R. Albert and A. Barabási, *Reviews of Modern Physics*, **74**, 47-97 (2002).
31. Barabási AL, Oltvai ZN, *Nat Rev Genet.*, **2**, 5 (2004).
32. D.J. Watts, S.H. Strogatz, *Nature*, **393**, **6684**, 440-442 (1998).
33. S. R. Collins, P. Kemmeren, X. C. Zhao, J. F. Greenblatt, F. Spencer, F. C. Holstege, J. S. Weissman, and N. J. Krogan, *Molecular and Cellular Proteomics*, **6(3)**, 439-450 (2007).
34. C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields and P. Bork, *Nature*, **417**, 399-403 (2002).
35. H. Yu *et al.*, *Science*, **322**, 104-110 (2008).
36. L. Giot *et al.*, *Science*, **302**, 1727-1736 (2003).
37. J. Yu, S. Pacifico, G. Liu, R. L. Finley Jr., *BMC Genomics*, **9**, 461 (2008).
38. C. Stark C, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, amd M. Tyers, *Nucleic Acids Res*, **34**(Database issue), D535-D539 (2006).
39. N. Simonis *et al.*, *Nature Methods*, **6**, 47-54 (2009).
40. J.-F. Rual *et al.*, *Nature*, **437**, 1173-1178 (2005).
41. Prasad et al., *Nucleic Acids Research*, **37**, D767-D772 (2009).
42. K. Venkatesan *et al.*, *Nature Methods*, **6**, 83-90 (2009).
43. M. Molloy and B. Reed, *Random Structures and Algorithms*, **6**, 161-180 (1995).