BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 4 2013, pages 483–491
doi:10.1093/bioinformatics/bts729

Systems biology

Graphlet-based measures are suitable for biological network comparison

Wayne Hayes[1], Kai Sun[2] and Nataša Pržulj[2,*]

[1]Department of Computer Science, University of California, Irvine CA 92697-3435, USA and [2]Department of Computing, Imperial College London, London SW7 2AZ, UK

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Large amounts of biological network data exist for many species. Analogous to sequence comparison, network comparison aims to provide biological insight. Graphlet-based methods are proving to be useful in this respect. Recently some doubt has arisen concerning the applicability of graphlet-based measures to low edge density networks—in particular that the methods are 'unstable'—and further that no existing network model matches the structure found in real biological networks.

Results: We demonstrate that it is the model networks themselves that are 'unstable' at low edge density and that graphlet-based measures correctly reflect this instability. Furthermore, while model network topology is unstable at low edge density, biological network topology is stable. In particular, one must distinguish between average density and local density. While model networks of low average edge densities also have low local edge density, that is not the case with protein–protein interaction (PPI) networks: real PPI networks have low average edge density, but high local edge densities, and hence, they (and thus graphlet-based measures) are stable on these networks. Finally, we use a recently devised non-parametric statistical test to demonstrate that PPI networks of many species are well-fit by several models not previously tested. In addition, we model several viral PPI networks for the first time and demonstrate an exceptionally good fit between the data and theoretical models.

Contact: natasha@imperial.ac.uk

Received on June 12, 2012; revised on October 19, 2012; accepted on December 20, 2012

1 INTRODUCTION

Networks are used to represent relationships in molecular biology, such as those between genes in gene regulatory networks and proteins in protein–protein interaction (PPI) networks. A PPI network models the physical bindings between types of proteins in a cell, where a node represents a type of protein and an undirected edge exists between two nodes if the corresponding two proteins can physically bind to each other. The network of all such interactions in a species has been dubbed the *interactome*, and understanding its structure is an integral step towards understanding cellular systems.

Recent advances in the experimental determination of PPI networks, such as yeast two-hybrid (Y2H) screening (Arabidopsis Interactome Mapping Consortium, 2011; Ito *et al*., 2000; Rual *et al*., 2005; Simonis, 2009; Stelzl *et al*., 2005; Uetz *et al*., 2000) and mass spectrometry (MS) of purified complexes (Gavin *et al*., 2002, 2006; Krogan *et al*., 2006; Rigaut *et al*., 1999) have provided an abundance of network data. As these PPI networks are large and complex, it is necessary to develop efficient and biologically meaningful algorithms for their analysis. Because exactly solving many graph analysis problems is computationally intractable, heuristic and approximate methods must be devised.

Systematic measures of structure (also called *topology*) of large networks based on *graphlets* (small induced subgraphs of large networks), have been introduced (Pržulj, 2007; Pržulj *et al*., 2004). One is based on comparing graphlet frequencies between two networks and is termed *relative graphlet frequency* (RGF) distance (Pržulj *et al*., 2004). The other is based on comparing *graphlet degree distributions* (GDDs) between two networks and the agreement in GDDs of two networks is measured (Pržulj, 2007). Both allow quantification of similarity between the topology of two data networks (e.g. between two species), or comparison of a data network with a theoretically derived model network. Fourteen PPI networks of eukaryotic organisms were compared with four different random network models and the fit between the PPI and model networks has been assessed (Pržulj, 2007; Pržulj *et al*., 2004).

The use of graphlet-based measures for network comparison has been questioned (Rito *et al*., 2010) owing to the measures being 'unstable' in regions of low edge density. Empirical distributions of these scores were calculated and a novel non-parametric test for assessing the statistical significance of the fit between real PPI networks and theoretical models was introduced. The study found that none of the PPI networks of yeast and human that were considered were well-fit by the three theoretical models that were tested.

After a discussion of methods, this article is designed to lead the reader through the following observations. First, that RGF distance and GDD agreement (GDDA) are not unstable measures for low edge density networks, but instead that they correctly detect the statistically high variance in network structure that is present in low-density model networks. Second, that this high variability [described as 'instability' by Rito *et al*. (2010)] is present in low-density model networks, but neither in real PPI networks nor in the model networks needed to correctly represent them. Third, that current PPI networks are denser than those studied by Rito *et al*. (2010) and are well outside the highly variable low-density region. Fourth, that the highly variable region shrinks with increasing network size and is of negligible

*To whom correspondence should be addressed.

© The Author 2013. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

size for real-world networks. Finally, we use the non-parametric test proposed by Rito *et al.* (2010) to demonstrate that PPI networks of many species are well-fit by several existing network models.

2 METHODS

2.1 Graphs and graphlets

A *network* or *graph* $G(V, E)$ consists of a set V of n nodes and a set $E \subseteq V \times V$ of m edges connecting them. Because the maximum number of edges is $\binom{n}{2} = n(n-1)/2$, we define the *edge density* of G as $m \Big/ \binom{n}{2}$. The number of edges touching a node is called its *degree*; the node at the other end of an edge is called a *neighbour*. The *degree distribution* of G, DD_G, is defined by $DD_G(k)$, which is the number of nodes in G having degree k.

A *subgraph* of G is a graph whose nodes and edges belong to G. A subgraph S of graph G is *induced* if S contains all edges that appear in G over the same subset of nodes. A *graphlet* is a small, connected and induced subgraph of a larger network (Pržulj *et al.*, 2004). Figure 1 depicts all 29 of the possible graphlets that have 2, 3, 4 and 5 nodes. Within each graphlet, some nodes are topologically identical to each other; such identical nodes are said to belong to the same *automorphism orbit*. Figure 1 shows the 73 different automorphism orbits among the 29 graphlets [see Pržulj (2007) for details].

Graphlets have been used to measure similarity of structure of PPI networks and network models (Pržulj, 2007; Pržulj *et al.*, 2004) resulting in the proposition of mechanistic models by which these networks have evolved (Pržulj and Higham, 2006; Pržulj *et al.*, 2010). Also, they have been used to align PPI networks of different species and transfer annotation to unannotated parts of these networks, as well as to reconstruct phylogeny from network alignment scores (Kuchaiev *et al.*, 2010; Kuchaiev and Pržulj, 2011; Milenković *et al.*, 2010a; Memišević and Pržulj, 2012). Furthermore, they have been used to measure the structure around proteins in the human PPI network that resulted in phenotypically validated predictions of new members of melanin production pathways purely from the PPI network topology (Ho *et al.*, 2010; Milenković *et al.*, 2010b).

2.2 PPI networks

We analyse eight PPI networks of four eukaryotes and 10 PPI networks of prokaryotes and viruses (Table 1): *Homo sapiens* (human), *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans* (nematode worm) and *Drosophila melanogaster* (fruit fly), and *Arabidopsis*

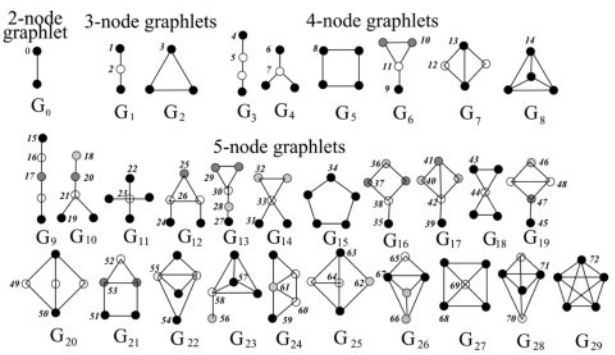


Fig. 1. Graphlets with 2–5 nodes, $G_0, G_1, \ldots, G_{29}$. The automorphism orbits are numbered from 0 to 72, and the nodes belonging to the same orbit are of the same shade within a graphlet (Pržulj, 2007)

(Arabidopsis Interactome Mapping Consortium, 2011). Here, HS is the human PPIs reported by Stelzl *et al.* (2005), and HG is the human PPIs reported by Rual *et al.* (2005); these two PPI networks were analysed both by Pržulj (2007) and Rito *et al.* (2010), and they are the first Y2H studies of the human interactome and hence are incomplete. HH is the human PPI network downloaded from HPRD version 9, released in April 2010 (Peri *et al.*, 2004). HR is the human PPI network collected by Radivojac *et al.* (2008). Finally, HB, WB, FB and YB are the PPI networks of human, worm, fruit fly and yeast, respectively, obtained from BioGRID version 3.1.74, released in March 2011 (Stark *et al.*, 2006). Viral PPI networks are from Fossum *et al.* (2009) and are described in Section 3.6. Bacterial PPI networks include CJJ (*Campylobacter jejuni*), ECL (*Escherichia coli*), SPP (*Synechocystis sp. PCC6803*) and MZL (*Mesorhizobium loti*). All self-loops, duplicate interactions and interspecies interactions were removed from the PPI data, as we consider only simple undirected graphs.

2.3 Random network models

We briefly review the random network models discussed in this article. An *Erdös–Rényi* (ER) random graph (Erdös and Rényi, 1959, 1960) is generated by fixing the number of nodes n in the network, and adding edges uniformly at random until a given density is reached. An ER-DD model is an ER model in which we force the degree distribution to be the same as that of a data network. *Scale-free* (SF) networks are those in which the degree distribution follows a power law (Barabási and Albert, 1999). SF-GD networks are scale-free networks based on a gene duplication model (Vaźquez *et al.*, 2003). If the degree distribution is the same as the data and the data are SF, we get a special case of an ER-DD model called random SF (SF-RND).

A *geometric graph* (Penrose, 2003) is one in which each node in the network represents a point in space and nodes are joined if their corresponding points are closer than some fixed global constant r. In this article, the geometric random networks (GEO) that we generate have their points distributed uniformly at random in Euclidean space. GEO-GD graphs are geometric graphs in which the points are distributed according to a *gene duplication* model (Pržulj *et al.*, 2010).

Table 1. PPI networks we analysed, ordered by size

PPI	Nodes	Edges	Density	Reference
HSV-1	47	100	0.09251	Fossum *et al.* (2009)
KSHV	50	115	0.09388	Fossum *et al.* (2009)
VZV	57	160	0.10025	Fossum *et al.* (2009)
EBV	60	208	0.11751	Fossum *et al.* (2009)
mCMV	111	393	0.06437	Fossum *et al.* (2009)
CJJ	1111	2988	0.00485	Parrish *et al.* (2007)
HS	1529	2667	0.00228	Stelzl *et al.* (2005)
MZL	1804	3094	0.00190	Shimoda *et al.* (2008)
HG	1873	3463	0.00198	Rual *et al.* (2005)
SPP	1920	3102	0.00168	Sato *et al.* (2007)
ECL	1941	3989	0.00212	Peregrín-Alvarez *et al.* (2009)
AT	2634	5529	0.00159	Arabidopsis Interactome Mapping Consortium (2011)
WB	2817	4527	0.00114	BioGRID (ver. 3.1.74)
YB	5607	57 143	0.00364	BioGRID (ver. 3.1.74)
FB	7372	24 063	0.00089	BioGRID (ver. 3.1.74)
HB	8920	35 386	0.00089	BioGRID (ver. 3.1.74)
HR	9141	41 456	0.00099	Radivojac *et al.* (2008)
HH	9465	37 039	0.00082	HPRD (ver. 9)

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Finally, the *STICKY* model is based on a stickiness index that is proportional to the degree of a node in a real network (Pržulj and Higham, 2006). It is based on the assumption that the higher the degrees of two nodes, the more likely they are to be neighbours.

2.3.1 The utility of model networks

> 'In the further development of science, we want more than just a formula. First we have an observation, then we have numbers that we measure, then we have a law which summarizes all the numbers. But the real glory of science is that we can find a way of thinking such that the law becomes evident.'—Richard P. Feynman, winner of the 1965 Nobel Prize in physics (Feynman, 1963).

We have been asked several times by biologists, including reviewers of our articles, why modelling biological networks is important and what models might teach us. It is easiest to answer this question by providing an example of another science in which models have become vitally important: physics.

Models provide us a way to think about a system in such a way that we can understand its behaviour; more importantly, a precise theoretical model allows us to *predict* a system's behaviour, so that we can learn how it reacts without having to experiment haphazardly on the system. For example, without a precise and predictive theory of gravity, we would not be able to understand the motion of planets, stars and galaxies, and NASA would not be able to launch robotic probes to the planets.

A first step towards understanding biological networks is to develop mathematical models of their structure; these models may not yet have predictive value, but they have descriptive value, in the same way that detailed descriptions of observations of movement of planets centuries ago provided data for Newton to develop a predictive theory of gravity. Biological network data today are noisy and incomplete, so the network structure may be hidden by noise and adversely filtered by sampling, population averaging and other biases in data collection, handling and interpretation (Collins *et al.*, 2007; Hakes *et al.*, 2008; Han *et al.*, 2005; Stumpf *et al.*, 2005; von Mering *et al.*, 2002). At this point in the development of structural models of biological networks, we cannot gain a detailed theoretical understanding of the networks. At best, we can try to measure statistical properties and develop models that have similar statistical properties to the real networks. Even this is an extremely difficult task. For one thing, graph theory tells us that large graphs can be almost infinitely complex in their structure; the number of undirected graphs with n nodes scales as $O(2^{n(n-1)/2})$, which is a function that grows almost unimaginably fast. For example, the number of graphs with 22 nodes is already comparable with the number of atoms in the observable universe ($\sim 10^{80}$); the number of graphs with 1000 nodes is $\sim 10^{150000}$.

Given these formidable difficulties, our goal at this juncture is not to find 'the' model that fits the data—that is hopeless. Instead our goal should be simply to aim the development of our models in the right direction.

2.4 GDD agreement

The notion of the degree has been generalized to *graphlet degree* (Pržulj, 2007). In particular, the standard *degree* of a node u counts how many edges it touches; in turn, the *graphlet degree* of u is a vector of 73 integers counting how many times u is centered on each of the automorphism orbits depicted in Figure 1. The *degree distribution* over degrees k of a graph is a distribution of how many nodes have degree k, i.e. touch k edges; in turn, a GDD is an analogous distribution for an automorphism orbit, e.g. the GDD corresponding to orbit 3 in Figure 1 measures how many nodes touch k triangles for each value of k. Hence, there are 73 GDDs for graphlets with up to five nodes; the first one of them is the degree distribution. More specifically, let G be a given graph. For each of the 73 automorphism orbits j shown in Figure 1, we denote by $d_G^j(k)$ the number of nodes in G that are touched k times by a corresponding graphlet at its orbit j. Because a dense graph tends to have a large graphlet degree for large values of k, we first downweight the graphlet degree by $1/k$ and then normalize, giving the normalized graphlet degree for automorphism orbit j,

$$N_G^j(k) = \frac{d_G^j(k)/k}{\sum_{l=1}^{\infty} d_G^j(l)/l} \quad (1)$$

For a particular automorphism orbit j, we can compare the orbit j degree distributions of two networks G and H by treating the normalized degree for orbit j as a vector in degree space (that has dimension equal to the largest degree). We compute the Euclidean distance between these two vectors as

$$D^j(G,H) = \frac{1}{\sqrt{2}}\left(\sum_{k=1}^{\infty}[N_G^j(k) - N_H^j(k)]^2\right)^{\frac{1}{2}} \quad (2)$$

where the $1/\sqrt{2}$ is a normalization constant that ensures the distance is always <1 (Przulj, 2010). Finally, we subtract each distance from 1 to get an 'agreement' instead of a distance, and sum over all the 73 values of j to get the GDDA

$$\mathrm{GDDA}(G,H) = \frac{1}{73}\sum_{j=0}^{72}(1 - D^j(G,H)) \quad (3)$$

If two networks G and H have a GDDA close to 1, then G and H have similar topology in the sense that their GDDs, scaled appropriately to their network size, are statistically similar.

3 RESULTS AND DISCUSSION

3.1 GDDA is sensitive, not unstable

A concern that 'GDDA has a pronounced dependency on the number of edges and vertices of the networks being considered' and that the 'GDDA score is not stable in the region of graph density relevant to current PPI networks' has been expressed (Rito *et al.*, 2010). Figure 2 illustrates the point: it depicts the dependency of GDDA on edge density when comparing two model networks with each other. The abrupt drop and raggedness of the GDDA curve at low densities is what was referred to as 'unstable GDDA'.

It is crucial to distinguish between a measure being unstable, versus a measure that correctly detects when the structure of a network is highly variable. We maintain that GDDA (and RGF distance, although space limitations preclude a discussion) faithfully detects genuine differences between the structure of various networks, thus making them of the latter sort. For example, consider two networks that have eight nodes and four edges each. The first consists of four isolated edges, while the other consists of one square and four isolated nodes. We can rightfully say that the two networks are different from each other, and both RGF distance and GDDA would detect this fact. Thus we see that it is not the measure that is unstable but the network structure itself. The measure simply reflects this 'instability' in networks with a relatively small number of edges.

Thus, although it is true that GDDA has an abrupt drop at low edge density, we believe that this drop is due to network properties having a statistically high variance at low edge densities. This high variance is evident in Figure 2, where the 1σ error bars are significantly larger at low densities than they are at high densities. This high variance in turn causes a low agreement when comparing two model networks that have low edge densities with each other. At higher edge densities, the statistical

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

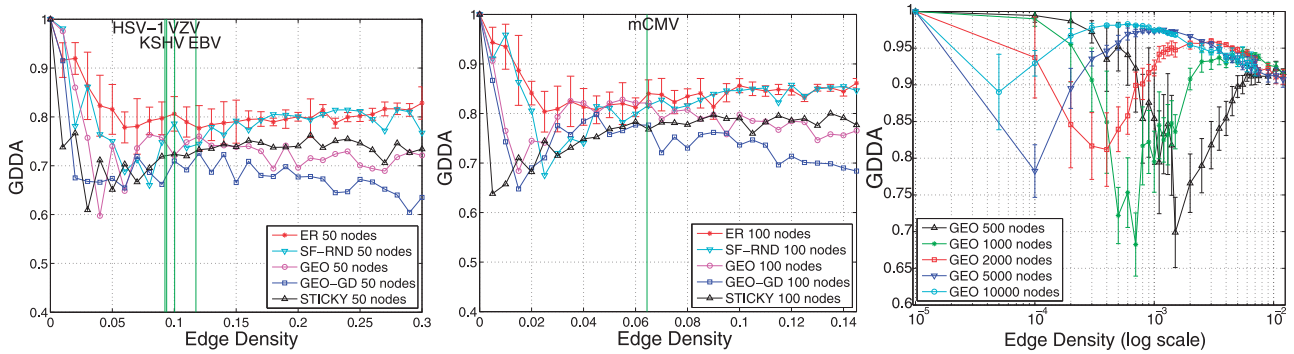


Fig. 2. GDDA versus edge density when comparing 30 random model networks with each other with (**a**) 50, (**b**) 100 and (**c**) 500–10 000 nodes. (Using >30 model networks resulted in virtually identical figures.) Panels (a) and (b) include one standard deviation error bars only for ER models, as all models have about the same variance as demonstrated in panel (c) for GEO networks. Green lines mark the edge densities of several viral PPI networks discussed in the text. In all cases, the viral PPI networks have densities outside the region of sensitivity

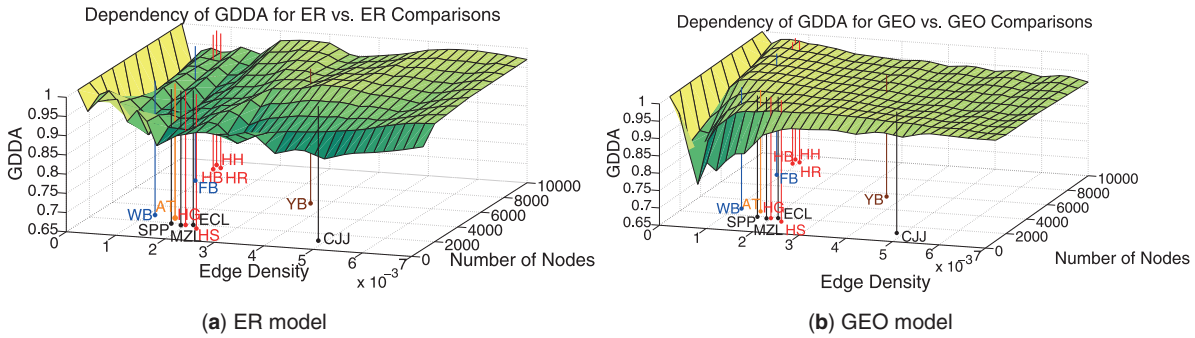


Fig. 3. The surface defined by 30×30 model-versus-model GDDAs as a function of number of nodes n and edge density ρ (empirically measured). (**a**) ER model, (**b**) GEO model. Each dot on the Cartesian plane represents the (n, ρ) of a real PPI network for the species from Table 1; they are grouped by color into kingdoms. As can be seen, in all cases the data networks are outside of the sensitive region for both ER and GEO models

properties of two model networks become more similar, resulting in a higher GDDA.

A simple analogy would be an experiment in which a fair coin is flipped n times in an attempt to experimentally determine the probability of heads. If $n = 0$, then both experiments will agree in whatever 'assumed' probability is assigned *a priori* (analogous to GDDA = 1 at zero edge density). If n is finite but small, then two experiments are likely to come up with substantially different estimates of the probability of heads, neither being close to the true value of 0.5. (This corresponds to the drop in GDDA at low but finite edge density.) However, as $n \to \infty$, the law of large numbers ensures that both experiments will provide similar estimates, both being close to 0.5. (This corresponds to the 'recovery' of GDDA as density increases.) Model networks differ from this analogy in that even as $n \to \infty$, the statistical difference between two models remains finite rather than approaching zero.

3.2 The unstable region is small in large networks, and natural networks have densities outside this region

We wish to quantify the extent of the 'unstable' region to decide if a given network has a density in the 'unstable' density region. The discussion associated with Figure 3 of Rito *et al.* (2010) makes it clear that the unstable region they refer to spans the density range [0, 0.01] for ER graphs with 500 nodes, and the range [0, 0.005] for GEO graphs with 500 nodes. Consider our Figure 2c. [The same analysis can be done on Figure 3b of Rito *et al.* (2010)]. It is clear that the minimum GDDA score for GEO graphs occurs at a density that is proportional to k_1/n for some value of k_1, where n is the number of nodes in the graph. Furthermore, the 'recovery' referred to by Rito *et al.* (2010) occurs at a density proportional to k_2/n for some value of k_2. This scaling is also apparent for all models if we observe Figure 2a and b: when going from panel (a) to (b) we double the number of nodes, simultaneously cut in half the extent of the horizontal axis and yet the two figures look similar. When discussing their Figure 3, Rito *et al.* (2010) say the unstable region for a 500 node GEO graph spans the density region $\rho \in [0, 0.005]$. From this we infer that $k_2 \approx 2.5$ for GEO graphs. A similar analysis of our Figures 2a and b and 3a of Rito *et al.* (2010) suggests $k_2 \approx 5$ for ER graphs. Thus, in general, it is reasonable to hypothesize that the unstable region spans the density region $\rho \in [0, k_2/n]$ for some value of k_2 that depends on the network model but is likely in the range $k_2 \in [2, 5]$. In any

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

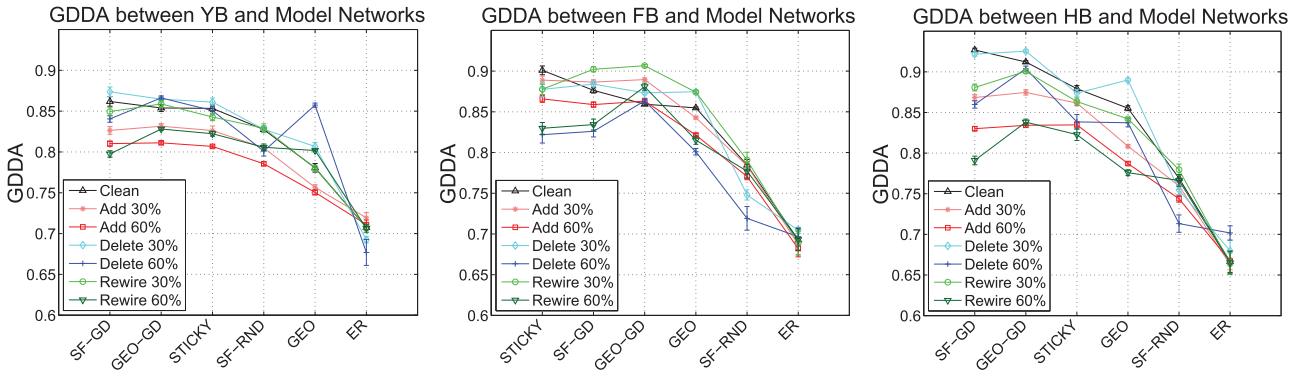


Fig. 4. GDDA is robust with respect to noise. In particular, we see that adding, deleting or rewiring as many as 60% of the edges in the model networks has little effect on which model best fits the data. Thus, conclusions about which models better fit the data are relatively insensitive to the fact that real-world network data contain substantial amounts of noise. Legend: YB = Yeast; FB = Fly; HB = Human

case, it is important to note that *the width of the unstable region in density space asymptotically shrinks with increasing n*.

Even small real-world networks, such as the viral ones that will be discussed later in reference to Figure 6, have densities outside the unstable region. In particular, Figure 2 depicts the density of viral networks (green vertical lines in Figure 2a and b) are in the 'stable' region, regardless of the assumed theoretical model.

Figure 3 demonstrates a similar effect for larger networks. We plot a surface depicting the dependency of GDDA on the number of nodes n and the density ρ for two types of model networks, ER and GEO. We also place each of the non-viral PPI networks from Table 1 in the (n, ρ) plane. As can be seen, none of these PPI networks are in the 'unstable' region of ER or GEO models. Note that the instability region and shape is different for different models; for example, the unstable region of GEO model is much smaller than the one of ER model with the same graph size and density. Furthermore, real PPI networks are not ER [e.g. Pržulj *et al.* (2004); Pržulj (2007); also Figure 6].

We note that GDDA can be sensitive to noise, whereas conclusions about network fit using GDDA are robust to noise. Figure 4 depicts how GDDA responds when we add, delete and rewire up to 60% of the edges in the model networks. We note that as noise is added, GDDA generally (though not always) decreases. However, the *ordering* of which model best fits the data undergoes only small changes when noise is added because all the curves generally have a negative slope, indicating that the ordering of models on the horizontal axis remains the correct ordering of model fit, independent of noise level. Noise has also been extensively discussed previously (Han *et al.*, 2005; Kuchaiev *et al.*, 2009; Stumpf *et al.*, 2005; Pržulj, 2007; Pržulj et al., 2004).

As can be seen in Table 1, the largest PPI networks have thousands of nodes and tens of thousands of edges. Assuming a value of $k_2 \in [2, 5]$, we see that most of the real-world networks in Table 1 have average densities higher than $4/n$, and all have average densities above $3/n$. Because natural networks also tend to have local densities higher than this [see our Section 3.4 as well as Colak *et al.* (2009); Pržulj *et al.* (2004); Gibson and Goldberg (2011)], they almost certainly are stable to GDDA measures.

We conclude that for large networks of substantial density, the width of the sensitive region is negligible, and that for even small networks that are relatively complete, their density is outside of the unstable region. This effect was not apparent in the studies performed by Rito *et al.* (2010) because (i) they neglected to look at fairly small but complete networks such as viral ones, (ii) they restricted the size of their model networks to just 2000 nodes and did not study how the unstable region shrinks with increasing n and (iii) they neglected to take into account that local densities may be higher (see Section 3.4 below).

3.3 Effect of more up-to-date PPI data

For yeast, Rito *et al.* (2010) analysed the earliest Y2H data (Ito *et al.*, 2000) with only ~800 nodes and edges, along with an early MS-based high confidence PPI network (von Mering *et al.*, 2002) with ~1000 nodes and 2500 edges. For human, they used the initial Y2H screens of Stelzl *et al.* (2005) and Rual *et al.* (2005) with only ~1700 and 3000 nodes and 3000 and 6700 edges, respectively. All these data are not only out-of-date now, it was out-of-date at the time Rito *et al.* (2010) did their study. It is of significantly lower density than more recently available networks, even networks that were available at the time of publication of *Rito et al.* (2010). They also used the current human PPI data from BioGRID, but split it up into two subnetworks, one with only PPIs from MS-based experiments and the other with only PPIs from Y2H experiments. By doing this, they decreased coverage and density of the currently available human PPI network, hence increasing false negative rate in each of the two subnetworks they constructed. Figure 5 depicts a timeline of data available from the year 2000 until the present, including data that were available significantly before the publication date of Rito *et al.* (2010) but not used by them.

3.4 Natural networks have high local densities in which GDDA is stable

We must distinguish between *average* and *local* density. Rito *et al.* (2010) used random model networks with the same *average* density of the data. However, data networks have *local* densities much higher than a uniformly dense random ER or GEO

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

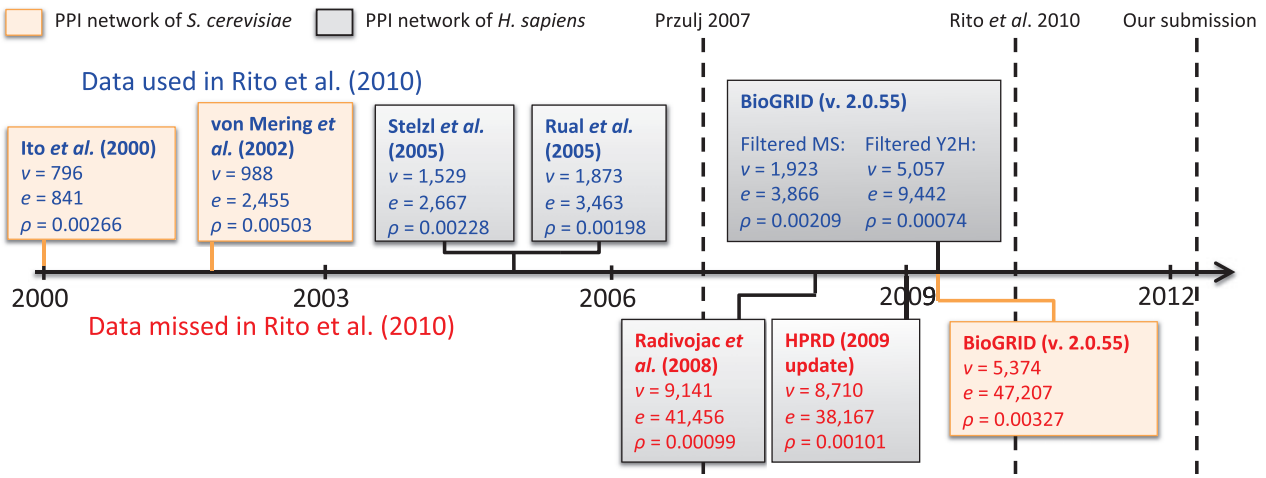


Fig. 5. A timeline of data available since 2000. Above the timeline are data used by Rito *et al.* (2010); below the timeline are data that could have been used by them but was not, resulting in lower edge densities in the networks they studied. Note that in reference to Biogrid v.2.0.55, Rito *et al.* (2010) filtered the PPI network into two networks by using the keywords 'Affinity Capture-MS' and 'Two-Hybrid', and they analysed the filtered networks separately instead of the whole network at once, thus artificially decreasing its density

Table 2. Graphlet frequencies in model and real PPI networks

Expected density at which first graphlet appears					Real graphlet frequencies of two human PPI networks	
Graphlet	ER 1000	ER 2000	GEO 1000	GEO 2000	HG ($n = 1873$, $\rho = 0.00198$)	HS ($n = 1529$, $\rho = 0.00228$)
G_1	0.00008	0.00003	~ 0.0011	~ 0.0004	76 807	36 776
G_3	0.00029	0.00011	~ 0.0033	~ 0.0013	675 350	296 346
G_4	0.00029	0.00011	"	"	2 018 299	435 650
G_8	0.01698	0.01070	"	"	20	1
G_9	0.00059	0.00025	~ 0.0060	~ 0.0025	9 056 316	3 438 523
G_{10}	0.00059	0.00025	"	"	32 102 612	7 462 899
G_{11}	0.00059	0.00025	"	"	54 733 928	5 067 616
G_{14}	0.00261	0.00130	"	"	1 433 133	118 173
G_{16}	0.00261	0.00130	"	"	1 466 355	740 181
G_{26}	0.02426	0.01573	"	"	409	13
G_{27}	0.02426	0.01573	"	"	103	9
G_{28}	0.03667	**0.02495**	"	**~0.0025**	**19**	0

To the left of the double vertical line, we reproduce parts of Tables 2 and 3 from (Rito *et al.*, 2010). These numbers represent the approximate edge density at which Rito *et al.* (2010) expect *one* graphlet of the type specified in column 1 to appear in a model network of either ER (columns 2 and 3) or GEO (columns 4 and 5) type with 1000 (columns 2 and 4) or 2000 (columns 3 and 5) nodes. To the right of the double bar are the actual frequencies of each graphlet type in human PPI networks HG (column 6) and HS (column 7), both of which have densities ρ of ~ 0.002. The bold numbers in row G_{28} are discussed in the text. Some graphlets are skipped to save space.

network would have. To see this, refer to our Table 2, where we reproduce part of Tables 2 and 3 from (Rito *et al.*, 2010). In it, we demonstrate that even the early human Y2H PPI networks, HG and HS, have many more large dense graphlets than one would expect in random network models with average densities of these PPI networks. For example, the highlighted numbers for graphlet G_{28} [taken from (Rito *et al.*, 2010)] indicate that we do not expect even one G_{28} to appear in a 2000-node ER network until the density reaches 0.02495—more than 10 times the density of the HG network—and yet the HG network, with 1873 nodes and a density of just 0.00198, already contains 19 copies of this highly dense 5-node graphlet. Even if the same human PPI network were modeled as a GEO network, it is expected to have zero copies of graphlet G_{28}, rather than 19. Furthermore, these human PPI networks contain *millions* of distinct 5-node graphlets, whereas the models predict only a few, at most. This tells us that a PPI network of low *average* density is not of *uniformly* low density, but instead contains highly dense subregions where the topology (and hence GDDA) is stable. This is also confirmed by a well-established observation that PPI networks have higher clustering coefficients than model networks (e.g. Pržulj *et al.*, 2004).

This effect is also seen in larger clusters on the human PPI network HS (taken from Stelzl *et al.*, 2005). We have used the

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Highly Connected Subgraphs algorithm (Hartuv and Shamir, 2000) to find dense subregions of HS network. The largest dense subregion contains >10% of the nodes (158 nodes) in the network and ~28% of its edges (744 edges); it has a local density of 0.06. For a network of $n = 158$ nodes, the worst-case density threshold marking the upper end of the instability region would be $5/n = 0.03$; a density of 0.06 for this largest dense subnetwork of HS is thus well above the threshold. We conclude the GDDA is stable when measuring the properties of this dense subnetwork.

3.5 Interpreting the results of non-parametric tests

RGF distance and GDDA provide a way to compare the statistical distribution of graphlets in two specific networks. Rito *et al.* (2010) provided a non-parametric test for assessing the fit between random model networks and real PPI networks. Given that our model networks are drawn at random from a particular distribution, each real PPI network will have a range of GDDAs with the random model networks. To obtain the distribution of GDDAs of model-to-model comparisons, we generate a number of model networks with the same size and density as the PPI network and compute the GDDAs across all pairs of these model networks. We also calculate the GDDAs between the PPI network and these random model networks to get the distribution of data-to-model comparisons. The difference between these two distributions can be used to evaluate the model fit. *Rito et al.* (2010) were correct in that, to determine if a particular PPI data network 'fits' a particular model, we need to look at the distribution of GDDAs of *many* data-to-model comparisons. If the data network is truly well-fit by the model, then we would expect the distribution of data-to-model GDDAs to be the same as the distribution of model-to-model GDDAs. This *non-parametric* test is a good test, in that given *any* measure of network similarity, it allows us to see if the data network is or is not statistically distinguishable from the model networks.

However, Rito *et al.* (2010) use the amount of overlap between these two distributions (model–model GDDA distribution versus model–data GDDA distribution) as a simple binary criterion for whether the data fits the model, without regard to the actual values of the GDDA. We believe that this interpretation is too restrictive. We must keep in mind that these models are extremely preliminary and purely empirical. When fitting such rudimentary models to such noisy data, we must maintain a larger perspective than a simple 'yes/no' determination can provide. To *improve* our models, we need instead to consider some form of 'distance' between the model and the data. For example, consider the following two cases: (A) Model A has model–model GDDA distributed normally $\sim 99 \pm 0.5\%$ and model–data GDDA normally distributed $\sim 97 \pm 0.5\%$; (B) Model B has model–model GDDA distributed normally $\sim 80 \pm 15\%$ and model–data GDDA normally distributed $\sim 70 \pm 15\%$. The 'overlap' test would suggest that case (B) is better than case (A) because case (B) would have substantial overlap between distributions while case (A) has none. However, a model–data GDDAof 97% [case (A)] is phenomenally good (assuming the model is derived independently of the data) compared with a GDDA of just 70% [case (B)]. The fact that Model A has no overlap with the data may be attributed, for example, to the model being somehow too narrowly defined; perhaps the addition of a small amount of noise (to mimic noise in the data) would bring the model–data agreement down to 97%, providing a perfect overlap and a high GDDA. It is not clear to us that overlap alone should be the sole criterion for judging how well a model fits the data.

3.6 Viral PPI networks are well-fit by models

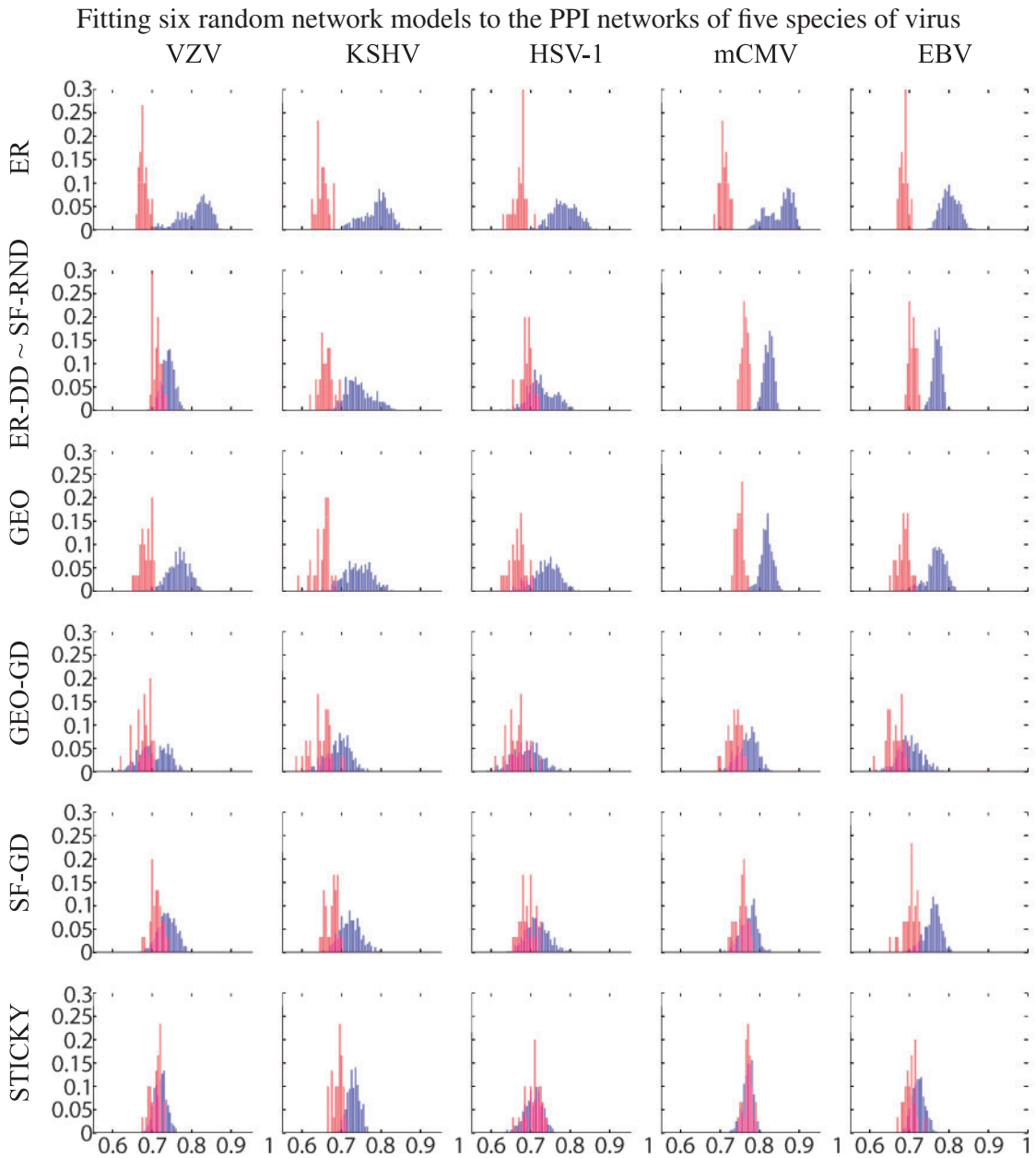
Large networks such as those from Table 1 are still noisy and *very* incomplete; the Arabidopsis Interactome Mapping Consortium (2011) claims only 2% coverage of the interactome (i.e. 98% false negative rate). In contrast, the substantially 'complete' networks of five herpes viruses depicted in Figure 6 (Fossum *et al.*, 2009) have 'only' ~50% false-negative rate (Sato *et al.*, 2007; Shimoda *et al.*, 2008). These five herpesvirus PPI networks are substantially complete in the sense that all possible pairs of proteins were tested for interaction using Y2H technology. The species are as follows: varicella-zoster virus (VZV), Kaposis sarcoma-associated herpes virus (KSHV), herpes simplex virus 1 (HSV-1), murine cytomegalovirus (mCMV) and Epstein–Barr virus (EBV). Figure 6 compares these complete PPI networks with six random graph models using the above-described non-parametric test (Rito *et al.*, 2010). In this test, we see that the ER random model is the worst fitting because in no case is there substantial overlap between the data–model distribution and the model–model distribution. As we move down the list of models in the figure, the amount of overlap generally increases until the STICKY model appears to have the greatest amount of overlap with four of the five viral PPI networks. To our knowledge, this is the first time that the structure of viral PPI networks has been modeled. Examining the biological reasons for the good fit of the STICKY model in 80% of the viral networks, and why it is a less good fit for KSHV, is a subject of future research.

Similar plots (not shown because of space limitations) of three bacterial PPI networks [MZL (Shimoda *et al.*, 2008), SPP (Sato *et al.*, 2007) and CJJ (Parrish *et al.*, 2007)], the functional interaction network of *E. coli* (Peregrín-Alvarez *et al.*, 2009), as well as the *Arabidopsis thaliana* PPI network (Arabidopsis Interactome Mapping Consortium, 2011), and PPI networks of yeast, worm, fly and human from BioGRID, all indicate that STICKY, SF-GD and GEO-GD (in that order) are the best fitting models for these networks.

4 CONCLUSIONS

In this article, we have examined the use of GDDA for biological network comparison. By generating the empirical distributions of GDDA scores, we have identified the edge density regions in which the topology of model networks is unstable. We show that the GDDA scores correctly measure this topological instability. These regions do not affect on the analysis of current PPI data, as current PPI data are dense enough to avoid these regions. We validate the use of GDDA for searching for well-fitting random models for PPI networks. We perform new fits to new PPI networks and demonstrate for the first time that five Viral PPI networks are will-fit by several models.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016



Overlap amounts (*i.e.*, shared area under the curve) for above histograms

Model/Species	VZV	KSHV	HSV-1	mCMV	EBV	Mean	StdDev
ER	0	0	0.0092	0	0	0.0018	0.004
ER-DD~SF-RND	0.2667	0.0299	0.369	0	0	0.1331	0.173
GEO	0.0161	0.0368	0.1782	0.0069	0.0529	0.0581	0.069
GEO-GD	0.3931	0.2839	0.5264	0.3586	0.3897	0.3903	0.088
SF-GD	0.4299	0.2046	0.5	0.5483	0.0966	0.3558	0.196
STICKY	0.5517	0.1138	0.7207	0.7414	0.4299	0.5115	0.256

Fig. 6. Comparing six random network models with the PPI networks of five species of herpesvirus (Fossum *et al.*, 2009) using Rito *et al.*'s non-parametric test. Each row represents one theoretical model, and each column represents one species. The horizontal axis is GDDA, the vertical axis is measured probability density. In each figure, the blue bars represent a histogram of GDDAs across all pairs of 30 randomly generated model networks with the same size and density as the PPI network of the corresponding virus; this gives us an expectation of how well the models compare with each other. The red bars represent a histogram of the GDDAs of the viral PPI network compared with the same 30 model networks. The quality of the fit is measured by the amount of overlap (i.e. shared area) under the distributions

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

We have demonstrated several network models (STICKY, SF-GD and GEO-GD) are good fits to the most complete existing viral PPI networks. We believe that this fact, in itself, is a significant milestone in the modelling of biological networks. The biological significance of the fits is not immediately clear, but the mere fact that we now have reasonably well-fitting models of viral PPI networks is a stepping stone towards greater understanding. Discerning how and why these models are not perfect, improving them and figuring out why STICKY is often but not always best, is a subject of future research.

ACKNOWLEDGEMENT

This work was supported by the European Research Council (ERC) Starting Independent Researcher Grant 278212, the National Science Foundation (NSF) Cyber-Enabled Discovery and Innovation (CDI) OIA-1028394, GlaxoSmithKline (GSK) Research & Development Ltd, and the Serbian Ministry of Education and Science Project III44006.

Conflict of Interest: none declared

REFERENCES

Arabidopsis Interactome Mapping Consortium (2011). Evidence for network evolution in an Arabidopsis interactome map. *Science*, **333**, 601–607.

Barabási,A.-L. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Colak,R. *et al.* (2009) Dense graphlet statistics of protein interaction and random networks. *Pac. Symp. Biocomput.*, **14**, 178–189.

Collins,S. *et al.* (2007) Toward a comprehensive atlas of the phyisical interactome of Saccharomyces cerevisiae. *Mol. Cell. Proteomics*, **6**, 439–450.

Erdös,P. and Rényi,A. (1959) On random graphs. *Publ. Math.*, **6**, 290–297.

Erdös,P. and Rényi,A. (1960) On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, **5**, 17–61.

Feynman,R.P. (1963) *Feynman Lectures on Physics*, Vol. 1. Addison-Wesley, Reading, MA.

Fossum,E. *et al.* (2009) Evolutionarily conserved herpesviral protein interaction networks. *PLoS Pathog.*, **5**, 13.

Gavin,A.C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Gavin,A.-C. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.

Gibson,T.A. and Goldberg,D.S. (2011) Improving evolutionary models of protein interaction networks. *Bioinformatics*, **27**, 376–382.

Hakes,L. *et al.* (2008) Protein-protein interaction networks and biology–what's the connection? *Nat. Biotechnol.*, **26**, 69–72.

Han,J.D. *et al.* (2005) Effect of sampling on topology predictions of protein-protein interaction networks. *Nat. Biotechnol.*, **23**, 839–844.

Hartuv,E. and Shamir,R. (2000) A clustering algorithm based on graph connectivity. *Inform. Process. Lett.*, **76**, 175–181.

Ho,H. *et al.* (2010) Protein interaction network topology uncovers melanogenesis regulatory network components within functional genomics datasets. *BMC Syst. Biol.*, **4**, 84.

Ito,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.

Krogan,N.J. *et al.* (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **440**, 637–643.

Kuchaiev,O. and Pržulj,N. (2011) Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, **27**, 1390–1396.

Kuchaiev,O. *et al.* (2009) Geometric de-noising of protein-protein interaction networks. *PLoS Comput. Biol.*, **5**, e1000454.

Kuchaiev,O. *et al.* (2010) Topological network alignment uncovers biological function and phylogeny. *J. R. Soc. Interface*, **7**, 1341–1354.

Memišević,V. and Pržulj,N. (2012) C-graal: common-neighbors-based global graph alignment of biological networks. *Integr. Biol.*, **4**, 734–743.

Milenković,T. *et al.* (2010a) Optimal network alignment with graphlet degree vectors. *Cancer Inform.*, **9**, 121–137.

Milenković,T. *et al.* (2010b) Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data. *J. R. Soc. Interface*, **44**, 353–350.

Parrish,J.R. *et al.* (2007) A proteome-wide protein interaction map for Campylobacter jejuni. *Genome Biol.*, **8**, R130.

Penrose,M. (2003) *Geometric Random Graphs*. Oxford University Press, Oxford, UK.

Peregrín-Alvarez,J.M. *et al.* (2009) The modular organization of protein interactions in Escherichia coli. *PLoS Comput. Biol.*, **5**, e1000523.

Peri,S. *et al.* (2004) Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*, **32** (Database issue), D497–D501.

Pržulj,N. (2007) Biological network comparison using graphlet degree distribution. *Bioinformatics*, **23**, e177–e183.

Pržulj,N. and Higham,D. (2006) Modelling protein-protein interaction networks via a stickiness index. *J. R. Soc. Interface*, **3**, 711–716.

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

Pržulj,N. *et al.* (2010) Geometric evolutionary dynamics of protein interaction networks. In: *Proceedings of the 2010 Pacific Symposium on Biocomputing (PSB)*. Big Island, Hawaii, January 4–8, 2010.

Przulj,N. (2010) Erratum to Biological network comparison using graphlet degree distribution. *Bioinformatics*, **26**, 853–854.

Radivojac,P. *et al.* (2008) An integrated approach to inferring gene-disease associations in humans. *Proteins*, **72**, 1030–1037.

Rigaut,G. *et al.* (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.*, **17**, 1030–1032.

Rito,T. *et al.* (2010) How threshold behaviour affects the use of subgraphs for network comparison. *Bioinformatics*, **26**, i611–i617.

Rual,J.-F. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.

Sato,S. *et al.* (2007) A large-scale protein–protein interaction analysis in Synechocystis sp PCC6803. *DNA Res.*, **14**, 207–216.

Shimoda,Y. *et al.* (2008) A large scale analysis of protein-protein interactions in the nitrogen-fixing bacterium Mesorhizobium loti. *DNA Res.*, **15**, 13–23.

Simonis,N. *et al.* (2009) Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. *Nat. Methods*, **6**, 47–54.

Stark,C. *et al.* (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.

Stelzl,U. *et al.* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.

Stumpf,M.P.H. *et al.* (2005) Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc. Natl. Acad. Sci. USA*, **102**, 4221–4224.

Uetz,P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.

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

von Mering,C. *et al.* (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016