BIOINFORMATICS

Vol. 30 ECCB 2014, pages i430–i437
doi:10.1093/bioinformatics/btu447

# Alignment-free protein interaction network comparison

Waqar Ali[1,*], Tiago Rito[1], Gesine Reinert[1], Fengzhu Sun[2] and Charlotte M. Deane[1]

[1]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK and [2]Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, CA 90089-2910, USA

## ABSTRACT

**Motivation:** Biological network comparison software largely relies on the concept of alignment where close matches between the nodes of two or more networks are sought. These node matches are based on sequence similarity and/or interaction patterns. However, because of the incomplete and error-prone datasets currently available, such methods have had limited success. Moreover, the results of network alignment are in general not amenable for distance-based evolutionary analysis of sets of networks. In this article, we describe Netdis, a topology-based distance measure between networks, which offers the possibility of network phylogeny reconstruction.

**Results:** We first demonstrate that Netdis is able to correctly separate different random graph model types independent of network size and density. The biological applicability of the method is then shown by its ability to build the correct phylogenetic tree of species based solely on the topology of current protein interaction networks. Our results provide new evidence that the topology of protein interaction networks contains information about evolutionary processes, despite the lack of conservation of individual interactions. As Netdis is applicable to all networks because of its speed and simplicity, we apply it to a large collection of biological and non-biological networks where it clusters diverse networks by type.

**Availability and implementation:** The source code of the program is freely available at http://www.stats.ox.ac.uk/research/proteins/resources.

**Contact:** w.ali@stats.ox.ac.uk

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

## 1 INTRODUCTION

The ability to recreate evolutionary relationships between biological objects has driven much of our increased biological understanding. In particular, the reconstruction of phylogenetic trees using sequence data is now commonplace and has provided evidence for evolutionary mechanisms such as mutation, insertion and deletion.

Similar to sequence data, over the past decade, the amount of available interaction data between proteins has been steadily increasing. These data are routinely represented as protein–protein interaction (PPI) networks, with proteins as nodes and interactions as edges. Despite the abundance of PPI data now available, there is no tree reconstruction method based purely on these networks. Yet, it is conjectured (Sharan and Ideker, 2006) that such trees based on networks may give rise to a step change in biology, much as tree-building methods from sequences did. There already are more elegant methods available for constructing species trees, e.g. based on genomic sequences. Trees built from PPI networks would span a limited species set and are by themselves perhaps not of interest. In contrast, the methodology that is able to correctly build a phylogenetic tree from interaction data would be much of interest, as it should reveal information about the evolutionary mechanisms at play in biological networks. Such a method would also be useful in other domains where alternative means of generating taxonomies are not available.

There is currently a lack of consensus regarding likelihood-based statistical models for PPI network evolution (Ratmann *et al.*, 2009). Moreover, often phylogenetic tree reconstruction methods based on distances perform better and are more robust to mis-specification than maximum-likelihood methods (Gonnet, 2012; Huelsenbeck and Hillis, 1993). Hence, we concentrate here on the development of a one-dimensional network comparison statistic, which can be used to compile a distance matrix to build trees.

The most tractable methods for network comparison are those which compare at the level of the entire network using statistics that describe global properties (e.g. Ratmann *et al.*, 2009), but these statistics are not sensitive enough to be able to reconstruct phylogeny or shed light on evolutionary processes. In contrast, there are several network alignment-based methods that compare networks using the properties of the individual proteins (nodes) e.g. local network similarity and/or protein functional or sequence similarity (Flannick *et al.*, 2009; Phan and Sternberg, 2012; Singh *et al.*, 2008). The aim of these methods is to identify matching proteins/nodes between networks and use these matching nodes to identify exact or close subnetwork matches. A few of these methods have been expanded to the multiple network problems (Flannick *et al.*, 2009; Liao *et al.*, 2009). These methods are usually computationally intensive and tend to yield an alignment that contains only a relatively small proportion of the network, although this has been alleviated to some extent in more recent methods (Alkan and Erten, 2014; Hu *et al.*, 2014; Patro and Kingsford, 2012).

The analysis is further confounded by a large number of false positives and false negatives thought to be present in current PPI data. Ali and Deane (2010) studied the effect of errors and incompleteness in alignments of simulated networks and estimated that only nearly complete networks ($>90\%$) can produce reliable alignments. Finally, PPI network alignment methods are all based on the loose premise that the respective orthologs of two interacting proteins also interact, forming pairs of so-called *interologs*, and/or that orthologs will share neighbourhood topology. While there is some evidence for the existence of such conserved interactions across species (Matthews *et al.*, 2001), particularly in proteins with high sequence similarity,

*To whom correspondence should be addressed.

Lewis *et al*. (2012) found, taking the noise and incompleteness of the data into account, that the fraction of correctly transferred interactions is at most 3%, between reasonably diverese species, even if orthologs are defined as those proteins matched with a blast $E_{val} \leq 10^{-10}$ (Altschul *et al*., 1997). Moreover, current evidence points to a far larger rate of change in PPIs than expected; specific interaction matches seem not to be the rule, but rather the exception (Lewis *et al*., 2012; Shou *et al*., 2011).

Thus, we do not follow the network alignment paradigm, but instead we take our lead from alignment-free sequence comparison methods that have been used to identify evolutionary relationships (Liu *et al*., 2011a). Alignment-free methods based on *k*-tuple counts (also called *k*-grams or *k*-words) have been applied to construct trees from sequence data (Song *et al*., 2013). A key feature is the standardization of the counts to separate the signal from the background noise. Inspired by alignment-free sequence comparison, we use subgraph counts instead of sequence homology or functional one-to-one matches to compare networks. Yet, even when comparing synthetic networks of the same size, generated from the same model, their small subgraph content can be volatile (Rito *et al*., 2010). Our proposed method, Netdis, compares the subgraph content not of the networks themselves but instead of the ensemble of all protein neighbourhoods (ego-networks) in each network, through an averaging many-to-many approach. The comparison between these ensembles is summarized in a Netdis value, which in turn is used as input for phylogenetic tree reconstruction.

The biological intuition for our new approach is based on a collection of results. The idea of using the subgraph content to build a distance between networks arises because motifs and modules have long been identified as important components of biological networks (Wagner *et al*., 2007; Zhu *et al*., 2007) and have been conjectured to play an important role in evolution (Liu *et al*., 2011b); see also Cootes *et al*. (2007) and Pržulj (2007) for previous explorations of biological network comparison based on subgraph counts. A key idea of our article is not to compare just the subgraph content of two networks, but instead the ensembles of subgraphs of the two networks, as hinted at in Rice *et al*. (2005). This use of subgraph content ensembles means that Netdis requires only the interaction data and is very different in concept from PPI network alignment methods. Netdis differs from standard subgraph count approaches in two key aspects: it introduces the ensemble view and applies a standardization that controls for background noise.

We first use our method on simulated networks using several random graph models and show that it correctly classifies networks by model type even when confounded by varying network density and size. We then use it to successfully reconstruct the correct phylogenetic tree for the set of organisms for which significant PPI data are currently available. Our ability to reconstruct correct phylogenies adds evidence that PPI data retain evolutionary information, despite the lack of conservation of individual interactions.

We have also investigated our method's ability to separate by type a large set of networks from several biological and non-biological domains. The resulting tree is found to be highly clustered by domain type, outperforming a recent and far more computationally intensive community-detection approach.

## 2 APPROACH

### 2.1 Neighbourhoods: ego-networks

Our method for network comparison is based on the core concept that similar networks will, on average, contain similar local neighbourhoods. We count the occurrence of subgraph shapes in the local neighbourhoods of all nodes in a network. This method was chosen rather than just counting the number of subgraphs in the entire network, as the latter will be strongly influenced by factors such as network size and density and, therefore, may be too coarse for many network comparisons.

Our notion of the neighbourhood of a protein is that of a two-step ego-network, also called ego-network of radius two (see, for example, Pattison, 1993). The two-step ego-network of a protein/node *p* is the (sub) network consisting of all nodes within two edges of *p*, also including all the edges between those nodes. In general, *r*-step ego networks could be used based on the application area. The radius should be chosen so that the set of ego-networks displays reasonable variability. Here we choose radius two, as the average shortest path length in protein interaction networks tend to be of the order four, and hence ego-networks with larger radii often contain a large proportion of the overall network. For very dense networks, ego-networks of radius one may be more appropriate.

### 2.2 An overview of the Netdis measure

Our algorithm starts by extracting for each query network the set of two-step ego-networks of all nodes. For each two-step ego-network, we count the number of occurrences of all 3 to 5-node induced subgraphs, or graphlets (Pržulj, 2007). The counting algorithm uses a combinatorial subgraph enumeration approach (Hoevar and Demar, 2014). Counting induced subgraphs on *k* vertices as opposed to any subgraphs on *k* vertices means that all the edges between the *k* vertices in the subgraph must be present in the larger graph, and absent edges in the subgraph must also be absent in the larger graph.

Every node in the network is hence associated with a *k*-nodes subgraph count vector ($k$ = 3, 4 or 5) for its corresponding two-step ego-network (see Fig. 1). One could also use $k = 6$ and above, but the counts of larger subgraphs can be extremely low in many networks and also computationally expensive to enumerate. Once the ego-network of a node has been associated with a count vector for all subgraphs contained in it, these counts are centred according to the size and density of the ego-network.

While ideally the counts would be centred using a suitable null model, currently no good probabilistic model for PPI networks is available that replicates the *k*-subgraph content of a network (Rito *et al*., 2012). Instead we use the counts from a gold-standard network as a proxy for the expected counts, where the size and the graph density of the size-two ego-network are taken into account.

For each *k*-node subgraph, we sum the centred counts of all two-step ego-networks in the query network. The final sum vectors have length 2 for $k = 3$, length 6 for $k = 4$ and length 21 for $k = 5$. These sum vectors are then used as input to a self-standardizing statistic, which we call $netd_2^S(k)$. In a final step, we calculate a symmetric matrix containing the $netd_2^S(k)$ values for pair of networks in the set. The resulting matrix can be used to

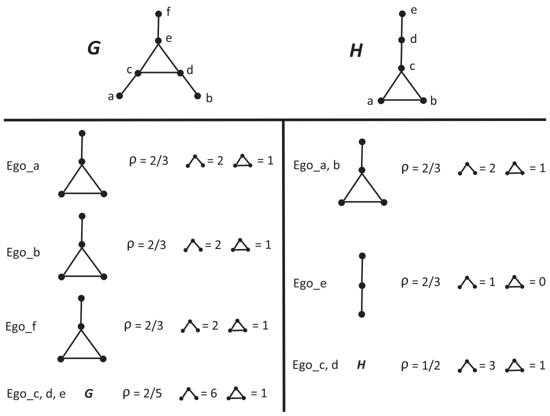
G
f
e
c
d
a
b
H
e
d
c
a
b
Ego_a ρ = 2/3 ∧ = 2 △ = 1
Ego_b ρ = 2/3 ∧ = 2 △ = 1
Ego_f ρ = 2/3 ∧ = 2 △ = 1
Ego_c, d, e G ρ = 2/5 ∧ = 6 △ = 1
Ego_a, b ρ = 2/3 ∧ = 2 △ = 1
Ego_e ρ = 2/3 ∧ = 1 △ = 0
Ego_c, d H ρ = 1/2 ∧ = 3 △ = 1

**Fig. 1.** Overview of the Netdis method on a pair of networks (*G*, *H*). Each network is associated with vectors of subgraph counts, calculated from all its two-step ego-networks (this figure shows counts for only subgraphs of three nodes). These count vectors, normalized by a background expectation, are then used to calculate the distance measure between the pair of input networks. See Supplementary Section S3 for a detailed calculation

cluster the candidate networks and the clustering represented by dendrograms.

## 2.3 Expectation of subgraph counts in neighbourhoods

In alignment-free sequence comparison, centering the counts by subtracting their means is crucial to avoid measuring background noise instead of signal (Reinert *et al*., 2009). For Netdis, we are faced with the problem that there is no good probabilistic model for subgraph counts in PPI networks available (Rito *et al*., 2012). Exploratory data analysis shows that the subgraph counts depend on the *graph density*, which is the fraction of observed edges over all potential edges.

To gauge the expected number of subgraph occurrences in an ego-network of a given graph density, we first create a histogram of the graph densities of all ego-networks in the gold-standard network. While the ideal comparison would be between ego-networks of the exact same density, there is not enough variability in the data for every possible density; hence, we use an adaptive binning approach. The ego-networks of the gold standard are binned according to density, where the number and size of bins is automatically adapted to ensure that each bin contains at least five samples. We then estimate the expectation per ego-network of each subgraph in a given graph density bin as follows.

Let $Q$ represent a gold-standard network with $q$ nodes and hence $q$ ego-networks. The ego-network of node $i$ contains, say, $n_i$ nodes, $e_i$ edges and has graph density $d_i = e_i \binom{n_i}{2}^{-1}$. For an ego-network of $Q$ with graph density $d_i$, we write $d_i \approx \rho$ if $d_i$ is in the graph density bin $\rho = \rho(Q)$. Let $w$ denote a particular induced subgraph with $k$ nodes and $N_{w,i}(Q)$ its number of occurrences in the ego-network of node $i$, with $n_i$ nodes, in $Q$. We scale these counts by the $\binom{n_i}{k}$ possible choices of $k$ nodes in the ego-network of node $i$. The average of the scaled $N_{w,i}(Q)$ for all $q$ ego-networks in the graph density bin is given by

$$E_w(Q,\rho) = \frac{1}{|\{i \in \{1,\ldots,q\} : d_i \approx \rho\}|} \sum_{\substack{i=1\ldots q:\\ d_i \approx \rho}} \frac{N_{w,i}(Q)}{\binom{n_i}{k}}.$$

For a query network $G$ and a given subgraph $w$ on $k$ nodes, we estimate the expected count of $w$ in the ego-network of node $i \in G$, with $d_i \approx \rho(Q)$ and $n_i$ edges, as

$$E_w^i(G,\rho) = \binom{n_i}{k} E_w(Q,\rho). \tag{1}$$

This estimated expected count serves as our background expectation.

Now, let $A(k)$ be the set of all $w$ subgraphs with $k$ number of nodes; here $k$ = 3, 4 or 5. For a particular induced subgraph $w$ with $k$ nodes, let $N_{w,i}(G)$ denote its number of occurrences in the ego-network of $i \in G$. We proceed as follows:

(1) For each node $i$ in $G$,

  (a) Build an ego-network of radius 2 around and including $i$.

  (b) Calculate for the ego-network of node $i$, the number of nodes $n_i$, graph density $d_i$ and the subgraph count vector, $N_{w,i}(G)$, for each $k$, with $k = 3$, 4 and 5.

  (c) If $d_i \approx \rho = \rho(i)$, calculate $E_w^i(G,\rho(i))$.

(2) Calculate,

$$S_w(G) = \sum_i (N_{w,i}(G) - E_w^i(G,\rho(i))). \tag{2}$$

To compare two networks, $G$ and $H$, we define three $netD_2^S(k)$ statistics by

$$netD_2^S(k) = \frac{1}{\sqrt{M(k)}} \sum_{w \in A(k)} \left( \frac{S_w(G) S_w(H)}{\sqrt{S_w(G)^2 + S_w(H)^2}} \right), k = 3, 4, 5,$$

where

$$M(k) = \sum_{w \in A(k)} \left( \frac{S_w(G)^2}{\sqrt{S_w(G)^2 + S_w(H)^2}} \right) \sum_{w \in A(k)} \left( \frac{S_w(H)^2}{\sqrt{S_w(G)^2 + S_w(H)^2}} \right)$$

is a normalizing constant so that $netD_2^S(k) \in [-1, 1]$ by the Cauchy–Schwarz inequality. The corresponding Netdis statistic is defined as,

$$netd_2^S(k) = \frac{1}{2}(1 - netD_2^S(k)) \in [0, 1]. \tag{3}$$

The pairwise Netdis values from Equation 3 are then used to build a distance matrix for all query networks. Note that three different distance matrices are defined, based on $k$ = 3, 4 or 5. We then use *Unweighted Pair Group Method with Arithmetic Mean* (UPGMA; Sokal and Michener, 1958) for building trees from the Netdis distance matrices. UPGMA is a heuristic greedy method that creates one cluster per network and sequentially merges the nearest pair of clusters by directly using the distance matrix until only two clusters remain. We ignore any branch lengths, as they would be difficult to interpret without a statistical model and before understanding the effect of errors. We used the *phangorn* package (Schliep, 2011) in R (R Core Team, 2013) to generate trees.

## 3 DATA

### 3.1 Synthetic networks from random graph models

We initially tested Netdis on simulated networks, namely, Erdös–Rényi (ER) random graphs (Erdös and Rényi, 1961), Erdös–Rényi graphs with fixed degree distribution (ERDD) (Newman, 2010), geometric random graphs (Penrose, 2003), geometric with gene duplication (Pržulj *et al.*, 2010), Chung–Lu model (Chung and Lu, 2002) and the duplication–divergence growth model (Middendorf *et al.*, 2005).

The particular ER model, which we use here, has $n$ labelled nodes connected by $m$ edges, which are randomly chosen from the $\frac{n(n-1)}{2}$ possible edges. The fixed distribution variant (ERDD) is constructed to have not just the same number of nodes and edges as a reference network, but also the same degree distribution. Geometric 3-dimensional random graphs (GEO3D) with parameter $r$ are constructed by assigning each node random coordinates in a 3-dimensional box of unit volume. Two nodes are connected by an edge if the Euclidean distance between them is at most $r$. A variant of this, incorporating biological intuition is the geometric with gene duplication model (GeoGD). The Chung–Lu (or Sticky) model constructs networks by assigning an index $\theta_i$ to every node $i$ for all $n$ nodes where $\theta_i = \frac{deg(i)}{\sqrt{\sum_{j=1}^{n} deg(j)}}$; here $deg(i)$ is the degree of node $i$. Each possible edge $i, j$ is then formed with probability $\theta_i\theta_j$. Finally, the duplication divergence model (DD) grows a network at each iteration $t$ by selecting a node random and duplicating it with all its edges. We use a variant of this model that specifies a symmetric node duplication model so that edges can be lost from both parent and child nodes on divergence.

### 3.2 Protein interaction data

To test Netdis on experimental data, we downloaded species-specific PPI data from the Database of Interacting Proteins (DIP; Salwinski *et al.*, 2004) and Human Protein Reference Database (HPRD; Keshava Prasad *et al.*, 2009). Only species having at least 500 physical interactions and $>15\%$ coverage were considered. Coverage is here a rough estimate of how many proteins have been probed for interactions given the expected proteome of the organism. We define it as a percentage by taking the number of nodes in the network divided by the estimated number of genes in the genome of the organism at hand. In total, we analyse five species: *Saccharomyces cerevisiae* (yeast), *Drosophila melanogaster* (fly), *Homo sapiens* (human), *Escherichia coli* (*E.coli*) and *Helicobacter pylori* (*H.pylori*). Human data came from HPRD (dated: September 2012) while the other four datasets were downloaded from DIP (dated: February 28, 2012). Table 1 summarizes the five PPI datasets used in the study.

**Table 1.** Network summaries for PPI data

| Species | Genes | Nodes | Edges | Coverage | $\rho^* \times 1000$ |
|---|---|---|---|---|---|
| Human | 21 224 | 9223 | 36 631 | 43.9 | 0.8 |
| Fly | 13 917 | 7565 | 22 800 | 54.3 | 0.8 |
| Yeast | 6692 | 5078 | 22 103 | 86.2 | 1.7 |
| *E.coli* | 4303 | 2968 | 11 604 | 68.9 | 2.6 |
| *H.pylori* | 1553 | 714 | 1,361 | 45.9 | 5.3 |

$^*\rho$—network density.

### 3.3 Networks from multiple disciplines

In a recent paper (Onnela *et al.*, 2012), the authors constructed a taxonomy of a large collection of networks obtained from a variety of sources. Their method is based on first probing the community structure within each network and then using summaries of the community structures to identify similar networks. The original dataset used by the authors contains 746 networks. We used all unweighted and undirected networks from this set, resulting in a total of 151 networks. These come from across the biological and social domains as well as model simulations (see Supplementary Section S3 for full details).

## 4 RESULTS

In all of the results that follow, unless stated otherwise, Netdis was calculated using $k = 4$ and the DIP core yeast interaction dataset (Deane *et al.*, 2002) as the gold standard. To test the robustness of the method, we also used a simulated gold-standard network using the ER model with 5000 nodes and 20 000 edges (see Supplementary Fig. S8 for results). While using Netdis with $k = 5$ is expected to increase the sensitivity of the method, preliminary analysis of the datasets used for this study indicated little benefit of the latter, at the cost of substantially more computing time spent in induced subgraph counting.

### 4.1 Netdis can separate different random graph model types

We simulated five networks for each of the six models described in Section 3.1, giving a total of 30 networks. The parameters for all of these simulated networks were chosen to have the number of nodes and edges match those of the DIP yeast network, although some models create self-loops and disconnected nodes, which lead to slight discrepancies. A pairwise distance matrix was then constructed between these networks and the resulting tree is shown in Figure 2. We observe a perfect clustering of the networks according to model type.

In the above analysis, all networks had the same or similar number of nodes and graph density. We also investigated the effect of varying network size and density on Netdis's ability to resolve the different models. For three models (ER, DD and Geo3D), we generated samples with either 1000 or 5000 nodes. We also varied the densities, using values of 0.003, 0.005 and

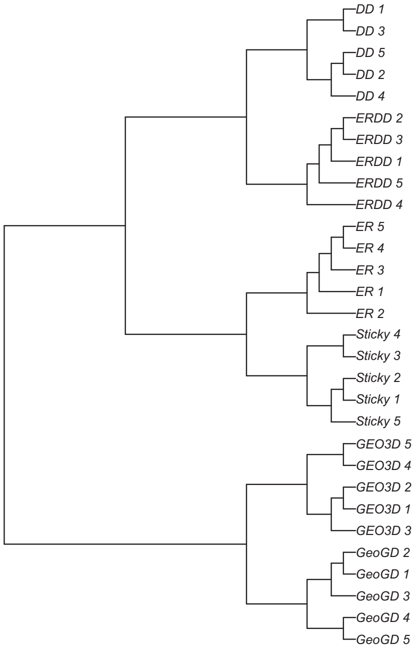

**Fig. 2.** Phylogenetic tree of simulated networks generated by Netdis. The method perfectly clusters together samples from the same model

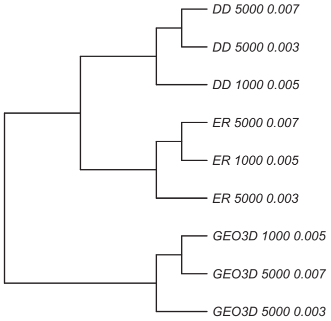

**Fig. 3.** Phylogenetic tree generated by Netdis for simulated networks of varying size and density. The model name is followed by number of nodes and network density

0.007. As shown in Figure 3, even with such varying sizes and densities, the correct clustering was achieved.

The simulated networks discussed so far are error-free. Introduction of error should make it harder to distinguish between networks from the different models. We introduced false negatives and false positives in the simulated networks (see Supplementary Section S2) and observed that for the simple case when all networks have the same size and density, Netdis can recover the correct grouping even with an error rate of 50% (Supplementary Fig. S11a). However, error has a larger adverse impact when the networks are already harder to cluster because of varying size and density (Supplementary Fig. S11b). The sensitivity of the method to errors is likely to be dependent on the dataset as well as error characteristics.

Our results on random graph models suggest that the different model types differ substantially from each other, and thus are perhaps not a rigorous test of the sensitivity of the method. This intuition was borne out when we reanalysed the dataset using modified versions of Netdis with the expectation in Equation 2 set to 0 (Netdis_ ne), or replacing the summation over all ego-networks in Equation 2 with just the subgraph count of the whole network (Netdis_ gc) and no correction for background expectation. Even these non-centred measures successfully generate the correct tree in all of the above cases (see Supplementary Figs S1 and S2).

Instead of ego-network subgraph counts, networks could also be grouped based on simpler properties such as the distribution of the node degrees or local clustering coefficients. To compare Netdis with such baseline measures, we defined Ddis and Cdis as the values of the Kolmogorov–Smirnov test statistic between the empirical distribution of nodes degrees and local clustering coefficients, respectively, for a given pair of networks. These values, serving as proxies for pairwise network distance were then used to generate phylogenetic trees for the simulated networks discussed above. The results (see Supplementary Figs S9 and S10) show that while these methods generate the correct clustering in the simplest case, they fail when the networks are variable in size and density, indicating the need for richer measures like Netdis in realistic settings.

While Netdis is completely different in approach to traditional network alignment, it is tempting to compare the results to such methods. Network alignment-based methods typically return results in the form of node or edge mappings and not a distance matrix for a given set of networks. However, one such method, MI-GRAAL (Kuchaiev and Pržulj, 2011) has been used in the literature to generate a phylogenetic tree of small viral PPI networks. MI-GRAAL is capable of solely topological network alignment and its *edge correctness score* was used as a proxy for network similarity measure to generate the distance matrix for the phylogenetic tree. We therefore tested MI-GRAAL on all of the above simulated datasets. For all simulated networks with 5000 or more nodes, we were unable to successfully finish alignment using MI-GRAAL. We therefore used only the networks of 1000 nodes and densities 0.003, 0.005 and 0.007 and the resulting tree is presented in Supplementary Figure S3. While the tree generated by MI-GRAAL is generally correct, two networks (ER_1000_0.003 and DD_1000_0.007) are wrongly clustered. The tree generated by Netdis for the same dataset, giving perfect clustering, is also given for comparison (Supplementary Fig. S4).

### 4.2 Phylogenies from protein interaction data

The currently accepted phylogeny between the species of Table 1 is depicted by the tree in Supplementary Figure S5. The tree is based on the NCBI taxonomy database (Sayers *et al.*, 2009), which incorporates a variety of phylogenetic resources including molecular and morphological characters. The tree generated by Netdis is depicted in Figure 4.

Out of the many possible rooted trees with five leaves, the correct clustering is obtained with fly next to human and yeast, and the two bacterial networks in a separate clade. This is despite the significant differences in number of nodes and edges, coverage and graph density. Moreover, Supplementary Figure S11c shows that even when introducing high levels of error in the PPI networks, the correct tree topology is reproduced by Netdis. The

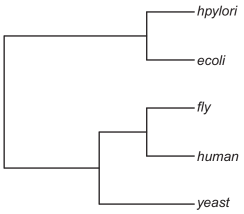

**Fig. 4.** Phylogenetic tree of PPI networks generated by Netdis

importance of the background expectation in Netdis becomes apparent when the tree is generated using Netdis_ ne instead (Supplementary Fig. S6a). In this case, the method does not recreate the correct phylogeny, and places fly next to *H.pylori*. The same is true when Netdis_ gc is used (Supplementary Fig. S6b). The simple network statistics based methods, Ddis and Cdis, also fail to generate the correct tree (Supplementary Figs S9c and S10c). When executing MI-GRAAL using only network topology on these data, the program failed during the alignment of the yeast and human network. We therefore only used the other four species (fly, yeast, *H.pylori* and *E.coli*) to generate the tree. In this case, MI-GRAAL recreates a generally correct tree with yeast and fly in a single clade (Supplementary Fig. S7a), but the bacterial species are split into two clades. The tree generated by Netdis for these four species is given in Supplementary Figure S7b. For PPI networks, as a baseline, one could also create phylogenies based on the number of orthologous interactions shared by a pair of networks. The results for such an approach are shown in Supplementary Figure S6c (see caption for method details). The tree is generally correct, although the approach is not applicable to other types of networks.

### 4.3 Classification of diverse empirical networks

As a systematic representation of data amenable to rigorous analysis and visualization, networks have become ubiquitous in recent years across many scientific and social disciplines. While these networks vary enormously in their detailed properties, methods that can group similar networks together could prove to be highly useful. We therefore compared the ability of Netdis with group networks by domain, against the method used by Onnela *et al.* on the data described in Section 3.3. As there is no agreed true dendogram available for these data, we first manually created a taxonomy by simply grouping the data based on type and assuming no further branching within groups. In total, we identify 13 groups, such as protein interaction, congressional voting, metabolic networks, ER random graphs, etc. Supplementary Figure S12 presents the complete manual taxonomy. We then created taxonomic trees for these networks using Netdis and Onnela *et al.*'s method. The resulting trees were split using the *cutree* function in R to give 13 groups (*cutree* can interpret a given tree as a cluster hierarchy that can then be merged/split using the underlying distance matrix to give the desired number of clusters). Finally, we compared the 13 groups from each of the methods with the manually created groups using the adjusted Rand index (RI) for cluster similarity (Hubert and Arabie, 1985). We also generated Monte-Carlo *P*-values for the similarity values by generating 50 000 samples from the null distribution as follows: each null sample was generated by creating a random tree topology with the same number of leaves as the dataset, and we then calculated the cluster similarity with the manual grouping. The *P*-value for an observation is then the fraction of null samples equal to or greater than the observed similarity. The best performing method is Netdis, with similarity index 0.011 (*P*-value: 0). Even without using background expectation (Netdis_ ne), a better grouping is achieved (RI: 0.01, *P*-value: $2 \times 10^{-5}$) than Onnela *et al.*'s method, which has a similarity index of 0.006 (*P*-value: 0.001). The clustered taxonomies generated by Onnela *et al.*'s method and Netdis are given in Supplementary Figures S13 and S14. Of particular interest is the fact that Netdis separates most of the metabolic and protein interaction networks into two distinct groups. The complete analysis for the 151 networks using Netdis consumed around 10 h of computing time on a standard desktop computer, while Onnela *et al.*'s method consumed 18 h on the same computer. The analysis could not be replicated using MI-GRAAL, as the program failed to provide alignment results for a large fraction of the dataset. We note that alignment-based methods are perhaps inherently ill-suited to the task of classifying large sets, which may include dense networks, as generating all possible pairwise network alignments is a computationally prohibitive task.

## 5 DISCUSSION

In this article, we add evidence to the idea that the topology of protein interaction networks alone contains evolutionary information without any additional biological data. Our results reveal that current PPI data are sufficiently abundant to derive correct phylogenetic relationships, at least between the model species. From PPI data with genome coverage of at least 15%, our method, Netdis, is able to deduce correct phylogenies between species without resorting to sequence homology.

We emphasize that Netdis is not proposed as a competitive method for the generation of phylogenetic trees for biological species; existing techniques based on molecular sequences already address this particular problem comprehensively. Our main result is that from the topology of protein interaction networks alone, it is possible to generate a correct phylogenetic tree. It is therefore worth considering the underlying assumptions of Netdis explore what they may reveal about protein interaction network evolution. The core principle of this work is that species that are more related will on average share more PPI network neighbourhoods that are topologically similar than unrelated species do. It is tempting to assign a biological interpretation to the Netdis algorithm: a given biological function is normally credited to a community of interacting proteins that act together to perform that function in the cell. Closely related species will have, on average, more of these communities in common. Our method detects this phenomenon by comparing ensembles of ego-networks derived from the PPI networks.

Netdis could also be used to consider the similarities between biological networks during a cellular or adaptive response, where it is thought that understanding the differences between these networks is key (Ideker and Krogan, 2012). Obtaining phylogenies of cell-states based on network data would provide a biological perspective on these phylogenies, which would be both new and completely free from sequence data. This perspective would further complement phylogenies derived from phenotypic traits with another type of molecular data. The method could be refined by adding information on the proteins, such as protein function, structure, subcellular location or age (Liu *et al*., 2011b; Rito *et al*., 2012); we would expect that including such information would increase the sensitivity of the method.

One particular focus of our ongoing and future research efforts will be the derivation of theoretically well-founded background expectations of subgraph counts used in Netdis. At present, the method relies on specifying a gold-standard dataset to estimate these expectations. While our results so far indicate that the method is robust to the choice of gold standard, it seems feasible that derivation of exact expectation formulas may increase method sensitivity and/or remove potential bias.

We conclude with noting that, as no other data besides network data are used as input to the method, many types of network data can be analysed together. This makes the method inherently different from the network alignment paradigm. As a proof of concept, we presented some results indicating its ability to correctly classify simulated networks from random graph models as well creating a reasonable taxonomy for a diverse mixture of empirical networks. The ability to highlight relationships between networks from different sources in a systematic way may help researchers across different fields to identify empirical analyses or theoretical models which are applicable to their specific problem.

## ACKNOWLEDGEMENT

The authors would like to thank Mason A. Porter for kindly providing the data from Onnela *et al*. (2012). The authors would also like to thank Robert Gaunt, Shuang Tian, Rebecca Walton and the anonymous reviewers, also of previous drafts of this article, for their valuable comments.

*Funding*: This work was supported in part by Fundação para a Ciência e a Tecnologia (FCT) through a PhD grant, and by the Biotechnology and Biological Sciences Research Council (BBSRC) and the Engineering and Physical Sciences Research Council (EPSRC) through the Systems Biology Doctoral Training Center (DTC) and the Oxford Center for Integrated Systems Biology (OCISB). GR, CD and WA acknowledge support from EPSRC grant EP/K032402/1.

*Conflicts of Interest*: none declared.

## REFERENCES

Ali,W. and Deane,C.M. (2010) Evolutionary analysis reveals low coverage as the major challenge for protein interaction network alignment. *Mol. Biosyst*., **6**, 2296–2304.

Alkan,F. and Erten,C. (2014) Beams: backbone extraction and merge strategy for the global many-to-many alignment of multiple PPI networks. *Bioinformatics*, **30**, 531–539.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Chung,F. and Lu,L. (2002) The average distances in random graphs with given expected degrees. *Proc. Natl Acad. Sci. USA*, **99**, 15879–15882.

Cootes,A.P. *et al*. (2007) The identification of similarities between biological networks: application to the metabolome and interactome. *J. Mol. Biol*., **369**, 1126–1139.

Deane,C.M. *et al*. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics*, **1**, 349–356.

Erdös,P. and Rényi,A. (1961) On the evolution of random graphs. *Bull. Inst. Internat. Statist*., **38**, 343–347.

Flannick,J. *et al*. (2009) Automatic parameter learning for multiple network alignment. *J. Comput. Biol*., **16**, 1001–1022.

Gonnet,G. (2012) Surprising results on phylogenetic tree building methods based on molecular sequences. *BMC Bioinformatics*, **13**, 148.

Hoevar,T. and Demar,J. (2014) A combinatorial approach to graphlet counting. *Bioinformatics*, **30**, 559–565.

Hu,J. *et al*. (2014) Netcoffee: a fast and accurate global alignment approach to identify functionally conserved proteins in multiple networks. *Bioinformatics*, **30**, 540–548.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif*., **2**, 193–218.

Huelsenbeck,J.P. and Hillis,D.M. (1993) Success of phylogenetic methods in the four-taxon case. *Syst. Biol*., **42**, 247–264.

Ideker,T. and Krogan,N.J. (2012) Differential network biology. *Mol. Systems Biol*., **8**, 565.

Keshava Prasad,T.S. *et al*. (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res*., **37**, D767–D772.

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

Lewis,A.C.F. *et al*. (2012) What evidence is there for the homology of protein-protein interactions? *PLoS Comput. Biol*., **8**, e1002645.

Liao,C.-S. *et al*. (2009) IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, **25**, i253–i258.

Liu,X. *et al*. (2011a) New powerful statistics for alignment-free sequence comparison under a pattern transfer model. *J. Theor. Biol*., **284**, 106–116.

Liu,Z. *et al*. (2011b) Evidence for the additions of clustered interacting nodes during the evolution of protein interaction networks from network motifs. *BMC Evol. Biol*., **11**, 133.

Matthews,L.R. *et al*. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res*., **11**, 2120–2126.

Middendorf,M. *et al*. (2005) Inferring network mechanisms: the drosophila melanogaster protein interaction network. *Proc. Natl Acad. Sci. USA*, **102**, 3192–3197.

Newman,M. (2010) *Networks: An Introduction*. 1st edn. Oxford University Press, USA.

Onnela,J.-P. *et al*. (2012) Taxonomies of networks from community structure. *Phys. Rev. E*, **86**, 036104.

Patro,R. and Kingsford,C. (2012) Global network alignment using multiscale spectral signatures. *Bioinformatics*, **28**, 3105–3114.

Pattison,P. (1993) *Algebraic Models for Social Networks. Structural Analysis in the Social Sciences*. Cambridge University Press, USA.

Penrose,M. (2003) *Random Geometric Graphs (Oxford Studies in Probability)*. Oxford University Press, USA.

Phan,H.T.T. and Sternberg,M.J.E. (2012) Pinalog: a novel approach to align protein interaction networks—implications for complex detection and function prediction. *Bioinformatics*, **28**, 1239–1245.

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

Pržulj,N. *et al*. (2010) *Geometric Evolutionary Dynamics of Protein Interaction Networks*. Chapter 20, pp 178–189.

R Core Team. (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ratmann,O. *et al*. (2009) From evidence to inference: probing the evolution of protein interaction networks. *HFSP J*., **3**, 290–306.

Reinert,G. *et al*. (2009) Alignment-free sequence comparison (I): statistics and power. *J. Comput. Biol*., **16**, 1615–1634.

Rice,J.J. *et al*. (2005) Lasting impressions: Motifs in protein-protein maps may provide footprints of evolutionary events. *Proc. Natl Acad. Sci. USA*, **102**, 3173–3174.

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

Rito,T. *et al*. (2012) The importance of age and high degree, in protein-protein interaction networks. *J. Comput. Biol*., **19**, 785–795.

Salwinski,L. *et al.* (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32** (Suppl. 1), D449–D451.

Sayers,E.W. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37** (Suppl. 1), D5–D15.

Schliep,K. (2011) phangorn: phylogenetic analysis in r. *Bioinformatics*, **27**, 592–593.

Sharan,R. and Ideker,T. (2006) Modeling cellular machinery through biological network comparison. *Nat. Biotechnol.*, **24**, 427–433.

Shou,C. *et al.* (2011) Measuring the evolutionary rewiring of biological networks. *PLoS Comput. Biol.*, **7**, e1001050.

Singh,R. *et al.* (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc. Natl Acad. Sci. USA*, **105**, 12763–12768.

Sokal,R.R. and Michener,C.D. (1958) A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.*, **28**, 1409–1438.

Song,K. *et al.* (2013) Alignment-free sequence comparison based on next-generation sequencing reads. *J. Comput. Biol.*, **20**, 64–79.

Wagner,G.P. *et al.* (2007) The road to modularity. *Nat. Rev. Genet.*, **8**, 921–931.

Zhu,X. *et al.* (2007) Getting connected: analysis and principles of biological networks. *Genes Dev.*, **21**, 1010–1024.