BIOINFORMATICS

Vol. 30 ISMB 2014, pages i237–i245
doi:10.1093/bioinformatics/btu283

Scale-space measures for graph topology link protein network architecture to function

Marc Hulsman, Christos Dimitrakopoulos† and Jeroen de Ridder*

Delft Bioinformatics Lab, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2628CD Delft, The Netherlands

ABSTRACT

Motivation: The network architecture of physical protein interactions is an important determinant for the molecular functions that are carried out within each cell. To study this relation, the network architecture can be characterized by graph topological characteristics such as shortest paths and network hubs. These characteristics have an important shortcoming: they do not take into account that interactions occur across different scales. This is important because some cellular functions may involve a single direct protein interaction (small scale), whereas others require more and/or indirect interactions, such as protein complexes (medium scale) and interactions between large modules of proteins (large scale).

Results: In this work, we derive generalized scale-aware versions of known graph topological measures based on diffusion kernels. We apply these to characterize the topology of networks across all scales simultaneously, generating a so-called graph topological scale-space. The comprehensive physical interaction network in yeast is used to show that scale-space based measures consistently give superior performance when distinguishing protein functional categories and three major types of functional interactions—genetic interaction, co-expression and perturbation interactions. Moreover, we demonstrate that graph topological scale spaces capture biologically meaningful features that provide new insights into the link between function and protein network architecture.

Availability and implementation: Matlab™ code to calculate the scale-aware topological measures (STMs) is available at http://bioinformatics.tudelft.nl/TSSA

Contact: j.deridder@tudelft.nl

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

1 INTRODUCTION

Understanding the living cell as a system of interconnected components is one of the key challenges of the post-genomic era (Kitano, 2002; Westerhoff and Palsson, 2004). To address this, many high-throughput screening techniques are used that chart the physical protein interaction network (PIN) of the living cell (e.g. Krogan *et al*., 2006; Ptacek *et al.,* 2005; Zhu *et al.,* 2009). Yeast, for instance, has one of the most comprehensive experimentally verified physical PINs (Stark *et al*., 2011). The interactions captured in these PINs entail the full range of chemical bonds between cellular components, ranging from protein complex formation and kinase signaling activity to transcription factor binding and DNA modifications.

Physical interactions give rise to functional interactions. These can be broadly categorized into (i) serial function interactions, such as the (causal) effects of a gene perturbation on downstream transcription, i.e. regulatory network interactions (Hughes *et al*., 2000), (ii) parallel function interactions, such as those arising from synthetic lethality (a genetic interaction; Costanzo *et al*., 2010; Phillips, 2008) and (iii) collaborative function interactions, such as co-expression in protein complexes.

Although some functional interactions are the result of a single physical interaction, most functional interactions arise as a result of a complex interplay between a collection of physical interactions. Consequently, analysis of the network architecture is important for understanding the functionality that it orchestrates (Barabási and Oltvai, 2004; Fuxman Bass *et al*., 2013). For this purpose, several graph topological measures have been proposed. These topological characterizations have, for instance, been used to uncover homology relations between proteins (Patro and Kingsford, 2012) or make predictions on protein function (Milenkoviæ and Pržulj, 2008). Moreover, it has been demonstrated that deletion of hub proteins—i.e. proteins with high network degree—is more often lethal than deletion of non-hub proteins (He and Zhang, 2006). This observation was explained by the observation that essential genes cluster in hub-enriched essential modules (Zotenko *et al*., 2008). Similarly, bottleneck proteins—i.e. proteins with high network betweenness—were found to be more essential and evolutionary conserved (Joy *et al*., 2005) and exhibit different expression dynamics than proteins with low betweenness (Yu *et al*., 2007).

Common graph topological measures, such as shortest-path length, Jaccard index, clustering coefficient and centrality measures, only capture topology in the direct vicinity of the node under investigation. They thus operate at a fixed 'zoom level', i.e. they do not take topological scale into account. This also holds for measures that characterize the centrality of a node with respect to the whole network. These include closeness centrality, betweenness centrality (Freeman, 1977), subgraph centrality (Estrada and Rodríguez-Velázquez, 2005), katz centrality (Katz, 1953) and eigenvector centrality (Bonacich, 1972).

The concept of topological scale is, however, important for studying how functional interaction emerges from the structure of the physical PIN (Fig. 1A). For instance, at the smallest scale, links in the PIN may directly implicate a functional relation between two proteins. At a medium scale, one could find functional relations that are characterized by a number of physical interactions, such as a small signaling cascade. In terms of graph

*To whom correspondence should be addressed.
†Present address: Department of Biosystems Science and Engineering, ETH Zurich, Basel; Swiss Institute of Bioinformatics, Switzerland

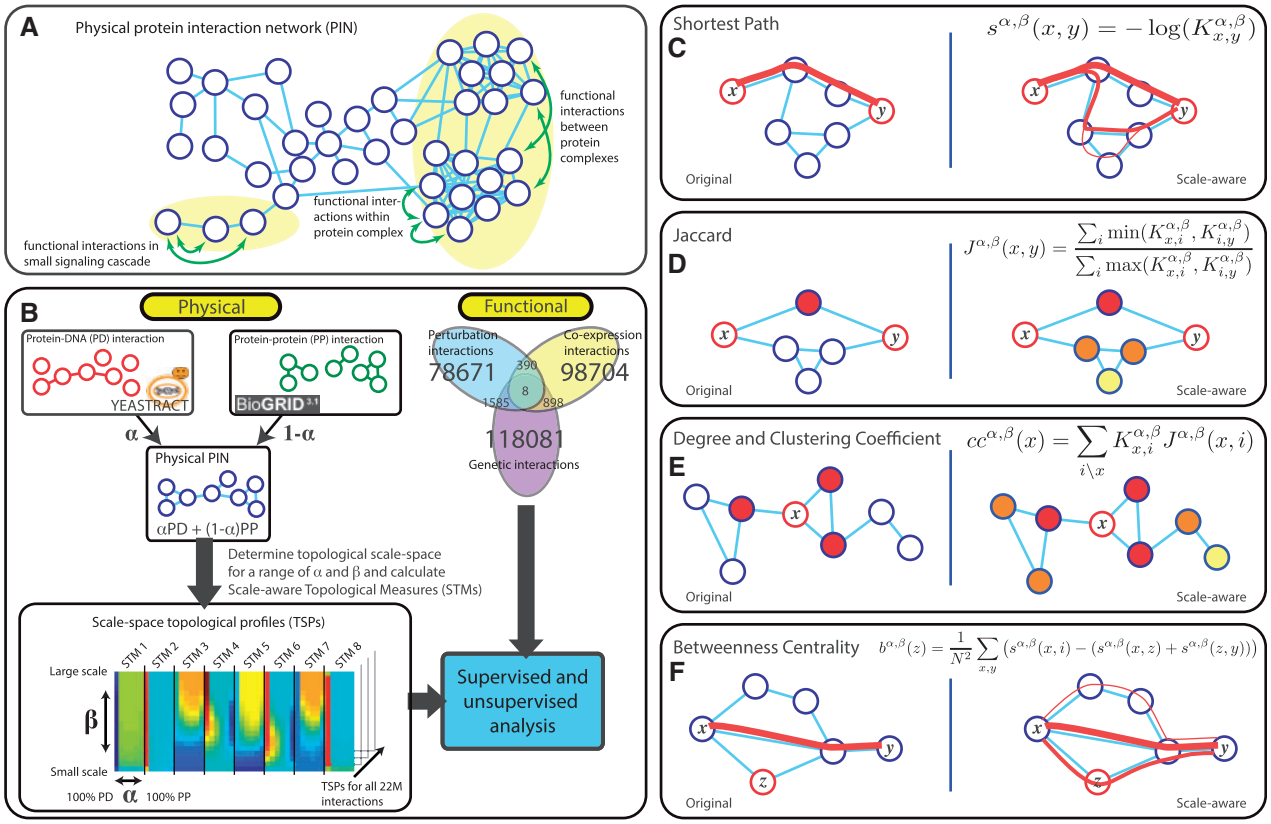

Fig. 1. (**A**) Schematic depection of a physical PIN in which topology dictates functional interactions. (**B**) Flowchart of data integration steps and application of the STMs. (**C**) Intuitive explanation of the shortest-path STM. For $\beta \to 0$, the STM behaves similar to the standard shortest-path measure. For higher values of β, it will start to take into account the number of shortest paths (bundle of paths) that connect x and y. Moreover, when β increases, paths that are longer than just the shortest path are considered. In this way, the overall network connectivity between two nodes is characterized. For $\beta \to \infty$, the measure will approach the logarithm of the number of nodes in the graph component containing x and y. (**D**) The Jaccard measure, which normally determines the fraction of common neighbors of two nodes, will take into account more extended neighborhoods with increasing scale. (**E**) Similar to the Jaccard, the centrality and clustering coefficient do not just take into account their direct neighbors but with increasing scale also the neighbors of neighbors. (**F**) The betweenness centrality measure (measuring the number of shortest path going through a specific node) is expanded so that it, with increasing scale, also takes into account longer paths. That is, although node z has a betweenness value of zero using the standard measure, the STM version still captures its role in linking x and y throughout a non-shortest path

topology, this could be captured by a small, linearly connected, graph structure. Finally, at larger scales, the network topology that describes complete protein complexes and pathways is important and should be captured to relate proteins functionally at a module level. The topology that describes this scale may consist of several densely connected subgraphs.

Because network topology occurs across a range of scales, traditional topological measures are inadequate to capture functional relations in the network. Instead, measures that include some kind of zoom parameter on topological scale are required. Measures that characterize 'meso-scale' topology have been proposed before (Jordán and Scheuring, 2002; Winterbach *et al.*, 2013b). Of particular interest is the scale-aware version of the subgraph centrality, proposed by Estrada, that resulted in superior ability to identify essential proteins (Estrada, 2010). However, the method used to incorporate scale does not extend to other topological measures and, moreover, was discrete in nature.

Here, we address this issue by introducing a scale-invariant description of the topology around or between proteins. These scale-aware topological measures (STMs) are built on the framework of diffusion kernels (Kondor and Lafferty, 2002), which can be seen as a network smoothing operation by means of diffusion. The level of smoothing determines the scale and can be tuned using a scale parameter. Application of kernel diffusion to a physical interaction network for a range of scales yields a graph topological scale-space. In spirit, such a scale-space is similar to image scale spaces used in computer vision applications. There, smoothing is used to obtain a family of derived images that describe the relevant image structure across all scales.

We derive a range of STMs that can be applied to this graph topological scale-space. STMs can be defined to characterize a single protein, such as the degree centrality STM, as well as to characterize the network connecting two proteins, such as the shortest-path STM. As our main focus is on functional interactions, we additionally derive link descriptors from the node-based STMs.

In the remainder of this article, we explore the use of STMs when applied to the comprehensive physical PIN of *Saccharomyces Cerevisiae* (yeast). Using supervised learning to predict three classes of functional interactions, i.e. genetic

interactions, perturbation interactions and co-expression interactions, we find that non-linear classifiers trained on STM features can reach area under the receiver-operating characteristic (ROC) curve (AUC) performances as high as 85%. Moreover, we demonstrate that clustering based on STM features reveals a pronounced substructure within each of the functional interaction classes that exhibit clearly distinct characteristics with biologically meaningful interpretations.

2 APPROACH

2.1 Constructing the physical PIN

Physical protein–protein (PP) interactions are obtained from BioGrid (Stark *et al.*, 2006) and Phosphogrid (Stark *et al.*, 2010) and collected in graph G_{PP} with adjacency matrix A_{PP}. Protein–DNA (PD) interactions are obtained from YEASTRACT (Teixeira *et al.*, 2006) and collected in graph G_{PD} with adjacency matrix A_{PD}. All interactions reported in each of these databases were included. Both A_{PP} and A_{PD} were made symmetric, thereby transforming directed interactions into undirected interactions. The resulting adjacency matrices contain 60 770 and 46 857 unique interactions (excluding self-interactions), respectively. To obtain the physical PIN, both graphs are combined in a linear fashion: $A^{\alpha}_{\mathrm{PIN}} = \alpha * A_{\mathrm{PD}} + (1-\alpha) * A_{\mathrm{PP}}$, where α is the mixing parameter.

2.2 Obtaining functional interactions

We derive three classes of functional interactions: co-expression, perturbation and genetic interactions. Throughout the text, these will be represented by symmetric binary adjacency matrices A_{coe}, A_{pt} and A_{gen}, respectively.

Co-expression interactions are calculated from the MegaYeast expression dataset (Gasch, 2012), which contains 501 yeast microarray experiments measuring expression during stress responses, sporulation and different cell cycle phases. A co-expression interaction between two proteins is counted if their encoding genes are among the top 100 000 correlating gene pairs (Pearson's $\rho > 0.67$).

Perturbation interactions are obtained from four datasets (Chua *et al.*, 2006; Hu *et al.*, 2007; Hughes *et al.*, 2000; van Wageningen *et al.*, 2010). A perturbation interaction is counted if one of their encoding genes exhibits a significant differential expression upon perturbation (knockout or overexpression) of the other. Differential expression is based on a *P*-value threshold of 0.01 as determined by the respective authors. A total of 80 654 perturbation-effect pairs are obtained, covering 622 perturbed genes. For the genes that were perturbed in different datasets, the resulting interactions are combined.

Finally, we obtained all genetic interactions reported by BioGrid (Stark *et al.*, 2006), totaling 120 580 interactions.

2.3 STMs and topological scale spaces

A wide range of graph topological measures exists (Fuxman Bass *et al.*, 2013; Winterbach *et al.*, 2013b). Here we focused on the following six measures: shortest-path length, Jaccard index, degree centrality, closeness centrality, betweenness centrality and clustering coefficient.

STMs are based on diffusion kernels (Kondor and Lafferty, 2002). The diffusion kernel function $k_\beta(A) = e^{\beta(A - deg(A))}$, with $deg(A)$ the degree matrix of A, applied to the adjacency matrix $A^{\alpha}_{\mathrm{PIN}}$ will result in the kernel matrix

$$K^{\alpha,\beta} = k_\beta(A^{\alpha}_{\mathrm{PIN}}) \tag{1}$$

The formulation of diffusion kernels ensures that $K^{\alpha,\beta} = (K^{\alpha,\beta})^T$ and $K^{\alpha,0} = I$.

One element in $K^{\alpha,\beta}_{x,y}$ gives the diffusion strength between node x and y and is a measure of their connectivity. Graph diffusion can be seen as network smoothing, as for increasing levels of diffusion β, the edge weights in the graph become more and more similar. Note that, in the continuous limit of a regular grid-based graph, taking the diffusion kernel corresponds to convolution with a Gaussian kernel (Kondor and Lafferty, 2002). Gaussian convolution is also used to construct scale spaces on images (Witkin, 1984). Importantly, for graphs, this smoothing process is dependent on the local topology and, moreover, can be used to describe it.

To construct a graph topological scale-space, we vary β across a range of values, which can be regarded as a scale parameter. $K^{\alpha,\beta}$ captures the evolution of the edge weights between all nodes across all scales. Using this representation, we can generalize the standard topological measures to work across different scale levels. In Section 5, we give a derivation for each STM, and in Figure 1C–F, we give an intuitive explanation for a few of them.

Some STMs characterize the topology around nodes (the clustering coefficient, centrality and betweenness STM), whereas other capture interactions within the network (the shortest-path and Jaccard STM). Node-based STMs can also be used to characterize network interactions by taking the average or difference between the scores at two connected nodes. In this fashion, we arrive at eight STMs for network interactions. Conversely, interaction-based STMs can also be converted to node-based STMs by taking the weighted average of the measures for all connections the node is involved in. Because the clustering coefficient is already defined as the weighted average of the Jaccard measure (see Section 5), this results in four unique STMs for nodes.

3 RESULTS

3.1 Supervised analysis

We evaluate the efficacy of the proposed STMs in a supervised setting, allowing us to compare them with the standard topological measures. To this end, we used the random neural network classifier (RNNC), combined with forward feature selection to determine the optimal set of features. A double-loop cross-validation was used to prevent selection bias and overtraining. Classifier performance was based on the AUC. More details can be found in Section 5.

3.1.1 STMs capture network structure useful for determining protein function To determine if STMs capture useful network structure that can contribute to the task of protein function prediction, we classified proteins to specific functional categories from the Munich Information Center for Protein Sequences (MIPS) catalog (Mewes *et al.*, 2004). Figure 2A reports the

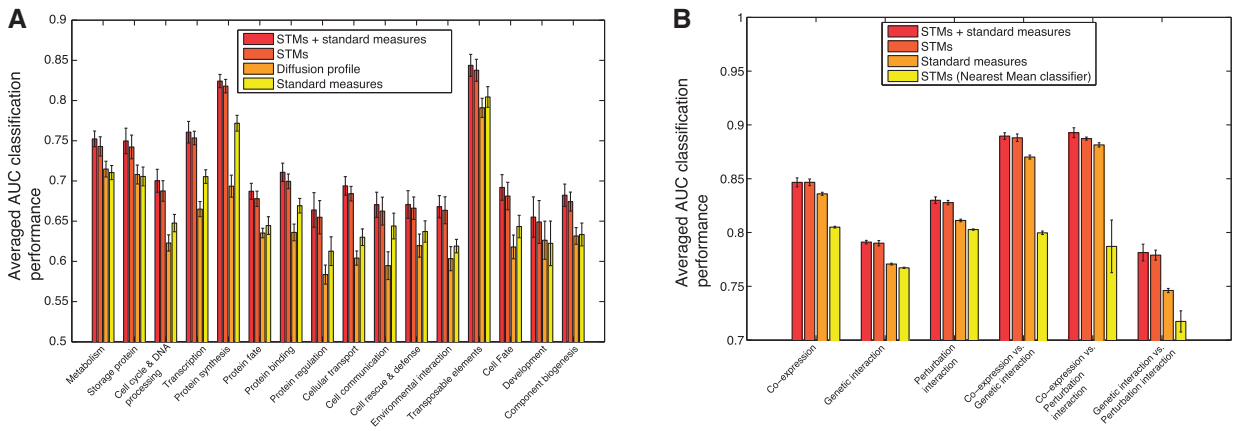

Fig. 2. (**A**) Classification performances for node-based STMs compared with kernel diffusion-based features and standard measures for 16 MIPS categories. (**B**) Classification performances for interaction-based STMs for three functional classes versus negative interactions and for pairs of the functional classes

one-versus-one AUC, averaged across all functional categories. It is apparent that STMs increase performance significantly. This is surprising, as others (Winterbach *et al*., 2013a) have noted that the fine details captured by more complex topological descriptors [e.g. graphlets Milenkoviæ and Pržulj (2008) and graph spectra Patro and Kingsford (2012)] do not significantly contribute to predictive performance, and that competitive performances could be obtained using simple topological measures.

Note that the poor performance obtained with a classifier trained on the diffusion kernel directly (middle bar) illustrates that the diffusion step alone is insufficient to capture the important topological structure in the network. Apparently, it is the combination of diffusion combined with a topological description that is important.

We reason that the observed performance gain is obtained because STMs incorporate information on the meso-scale topology, instead of adding additional fine-grained descriptions of the local topology. As a result, they describe how the local topology is embedded in the overall network. The improved classification suggests that this higher-level description of the network topology contains useful information for the task of protein function prediction.

3.1.2 Three classes of functional interaction can be distinguished using physical network topology Three main classes of functional interactions (co-expression, perturbation and genetic interactions) have extensive measurement coverage in yeast. We evaluated if these functional (positive) interactions could be distinguished from an equal number of randomly selected (negative) gene pairs for which there is no evidence of functional interaction. For all three datasets, we derived the previously described set of eight interaction-based STMs and used the RNNC combined with forward feature selection to discriminate them from the negative interactions.

The classification performances are summarized in the first three groups of four bars in Figure 2B. These results show that the three classes of functional interactions are accurately captured by the topological descriptions of the physical PIN, with a marked performance improvement for the STMs. This confirms earlier results that show that co-expression is linked to protein co-membership (Jansen *et al*., 2002) and that genetic interactions are linked to between-pathway and within-pathway PP interaction signatures (Kelley and Ideker, 2005). Moreover, it corroborates the observation that perturbation effects are informative in predicting the activation and inhibiting characteristics of PP and PD interactions (Ourfali *et al*., 2007).

The last three groups of four bars in Figure 2B depict the results when the classifier was used to discriminate between a combination of two functional classes. We found that the classes could be distinguished with surprisingly good performance, in the order of 0.9 AUC. This suggests that the interaction classes have different realizations in the underlying physical topology, and that the topological characteristics are accurately captured by the STMs. This may also explain why different classes of functional interactions have little overlap (Fig. 1B). For instance, the 80 654 perturbation and 100 000 co-expression interactions have only 390 interactions in common, which is just slightly more than one would expect by chance (375). Moreover, co-expression and perturbation interactions, which have a larger overlap, were easier to distinguish than genetic interactions and perturbation interactions.

Combining STMs with standard measures, results in only minor increase in performance for both protein function prediction (0.0082 on average) and functional interaction prediction (0.0021 on average). At the same time, performance is severely reduced if a simple linear classifier, the nearest mean classifier, is used. This indicates that STMs are able to capture most of the relevant network structure that can also be obtained with standard measures but need an advanced classifier to do so.

The learning curves and classification performance for the individual STMs are included in Supplementary Figure S1. The learning curve shows that the classifier requires more than a single topological descriptor to attain the best performance. On average, 23 topological descriptors were required, ranging across different STMs as well as different α and β levels, to attain maximum performance. Close-to-optimal performance was already reached using five topological descriptors. Combining STMs

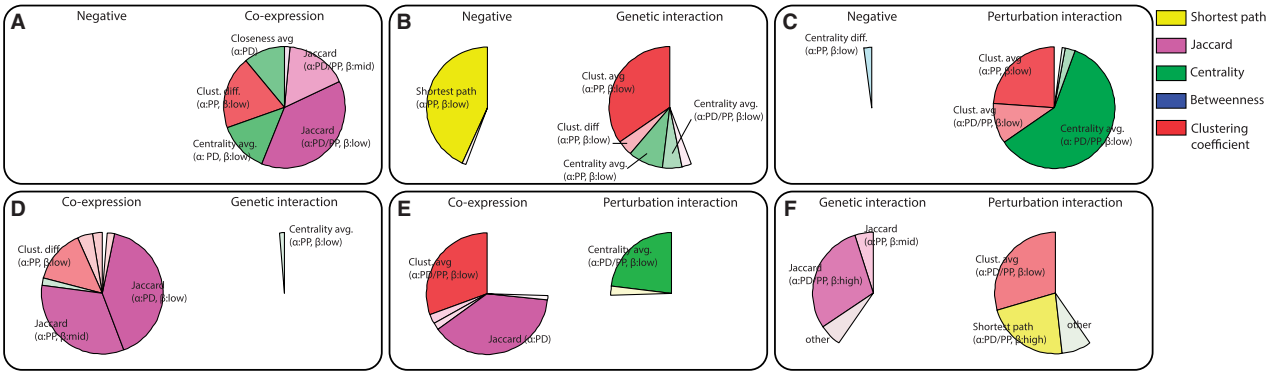

Fig. 3. (**A–F**) Top topological descriptors for each classification task, based on forward feature selection performed on five non-overlapping datasets. Large pie parts with intense colors represent the more relevant descriptors. Pie parts are assigned to classes based on the sign of their rank correlation with the class labels. Descriptors were grouped into a low- ($\beta < 0.2$), mid- or a high- ($\beta > 1$) scale group, as well as a PP ($\alpha = 0$), PD ($\alpha = 1$) or PD/PP mixture group. Standard (non-STM) topological features were also available to the feature selection and are represented without a β value. More in detail, the size of each pie part indicates the number of times a certain group of descriptors is represented in the top five descriptors, times their average rank value (where rank 1 corresponds to the value 1.0 and rank 5 to the value 0.2). Color indicates the STM type, while the intensity represents (again) the average rank value

outperforms the single STMs, indicating that the STMs augment each other, albeit to a small extent. Apparently, STMs can substitute each other to some degree. This may be a sign that fine details in the topology, of which different STMs capture different aspects, are not as important as the larger-scale topological structure, which is captured by every STM.

3.1.3 Topological signatures of functional interactions To determine which descriptors are the most important for each classification task, we performed feature selection. The robustness of the selected features was investigated by dividing the datasets in five non-overlapping subsets, and performing feature selection in each of them, ensuring five completely independent feature selections. Surprisingly, despite this, the first topological descriptor (a combination of STM type and α and β values) chosen in the selection was always the same for all but the genetic versus perturbation interaction task (where three of the five chosen descriptors were the same). Moreover, the second chosen descriptor was consistent in almost all cases except for some variation in α and β. This remarkably robust feature selection is a sign that each functional interaction class has its own unique topological signature.

A visual overview of the selected features, and thus the topological signatures, is given in Figure 3. Some of the signatures have interesting interpretations. In Supplementary Figure S2, the single descriptor (univariate) AUC performances are given.

For co-expression interactions, the first two selected measures are always the Jaccard and clustering coefficient STMs. The Jaccard measure is also selected at the medium scale further down the ranked list of selected descriptors, as well as for the classification experiments where co-expression interactions are contrasted to genetic and perturbation interactions (Fig. 3D and E). An explanation for this is that the Jaccard describes to what extent two proteins have the same position with respect to the surrounding topology. From Supplementary Figure S2, it can be seen that co-expression interactions are best predicted by low-scale STMs and depend only weakly on the mixture level of the PP/PD networks.

Genetic interactions are best characterized by the shortest-path and clustering coefficient STMs. Both results in Figure 3 and Supplementary Figure S2 show a preference for somewhat higher scales and a strong preference for PP interactions. The shortest-path measure is associated with the negative class, indicating that genetic interactions are represented by relatively short paths. The chosen scale is just below the threshold of the low β class with $\beta = 0.18$ (for all five independent repeats).

The second descriptor selected for genetic interactions (clustering coefficient STM) corroborates this picture. It is an indication that proteins with a genetic interaction are generally also embedded in a well-connected neighborhood. Taken together, this suggests that proteins that take part in a genetic interaction are well connected through the PP network, for instance, because they are part of a common complex, pathway or module but without necessarily having common regulators (PD interactions).

Finally, for perturbation interactions, we see a large focus on centrality measures, which are typical for regulators with many outgoing connections. Note that experimenter bias could play a role here, as perturbation experiments are typically performed for genes with a known regulatory role. On the other hand, the selection of this measure can also indicate that perturbations in more central proteins will induce more effects throughout the network and therefore have more perturbation interactions.

3.2 Unsupervised analysis

Next, we asked if the three broadly defined functional interaction classes each have a single descriptive topological structure in the physical network, or if multiple distinct topological implementations of the functional relations still could be present.

To address this question, we first mapped all interactions to a position in 2D space (Fig. 4), such that the differences in the topological characterization of these interactions were represented by their mutual distance in this space. The embedding reveals a clear structuring into distinct subgroups. Interestingly, interactions from the three interaction classes are not evenly distributed across the 2D map (Supplementary Fig. S3).

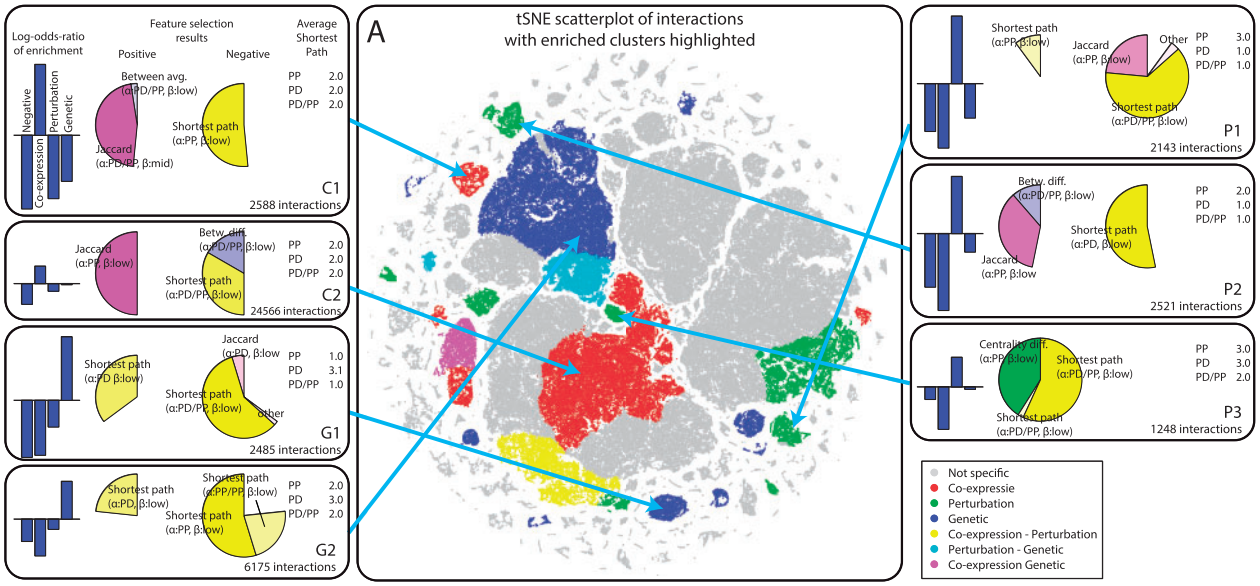

Fig. 4. (**A**) Representation of the interactions from the three functional interaction classes augmented with 100 000 negative interactions. The STM representation of each interaction (which constitutes a 960 dimensional space) was reduced using multidimensional scaling (t-SNE) to two dimensions. The t-SNE is an embedding technique, which aims to capture local structure in the data (Van der Maaten and Hinton, 2008). Interactions were clustered, and significant enrichments (Binomial test, Bonferroni-corrected $P < 0.01$ and a log odds ratio of > 1) are indicated by colors. (**C1–C2, G1–G2, P1–P3**) Detailed description of several clusters. The most significant cluster within each of the functional classes is shown, augmented with a few contrasting and/or supporting examples. Feature selection results are determined by contrasting the cluster to a random selection of interactions from other clusters five times the size of the cluster, which is repeated five times. The top three descriptors are visualized in a similar fashion, as shown in Figure 3. Visualizations of all other significant clusters can be found in Supplementary Figures S4–S6

To investigate this further, the data were clustered using a combination of spectral and hierarchical clustering. In this fashion, the 400 000 interactions were grouped into 498 clusters, of which 88 contained more than 500 interactions. Of these, 34 clusters (covering 40% of interactions) were significantly enriched (Fig. 4). For nine clusters, more than 75% of interactions in the cluster were of a single class, with one cluster (C1) containing 2523 interactions from the co-expression class (97%).

Next, feature selection was performed to investigate which topological descriptors best capture the topological structure of each cluster. To this end, interactions from each cluster were contrasted with a randomly selected set of interactions from the other clusters, five times the size of the cluster itself.

Some differences with the supervised analysis are expected, as the contrasting set of interactions contains interactions from all functional classes as well as the negative class in approximately equal balance. However, the results still reveal that, within each class of functional interactions, topological characterizations are shared among many of the enriched clusters (Supplementary Figures S4–S6). For instance, clusters enriched for the co-expression class are predominantly characterized by the Jaccard and shortest-path STMs at low scales (Fig. 4C1–C2), whereas clusters from the genetic class commonly have a shortest-path STM for the PP network and a negatively associated shortest-path STM (longest path) for the PD network (Fig. 4G1–G2). This indicates that, even though each class consists of interactions from a number of topologically distinct clusters, the topological characterization of the different types of functional interactions is still remarkably robust, confirming the picture painted in the classification experiment, as reported in Figure 3.

There are, however, still notable differences between the clusters. Clusters C1 and C2, for instance, differ in such a way that in Cluster C1, the Jaccard STM operates on the combined network at medium scale, whereas in Cluster C2, the PP network is selected at low scale. Both clusters contain interactions with a minimal shortest-path length of 2, indicating that they are not embedded in complexes.

For the perturbation interaction class, Clusters P1, P2 and P3 all depend on the shortest-path STM. However, Clusters P1 and P2 select it as a negative feature, indicating that short paths are preferred for the positive class. It is likely that these correspond to direct relations between a transcription factor and their targets. This is supported by the observation that all interactions in these clusters are directly connected in the PD network (the average shortest-path length in the PD network is 1). For Cluster P3, on the other hand, high shortest paths are favored, combined with a high centrality (which was also observed in the supervised analysis). This may correspond to propagation of the perturbation effect from a central regulator through a combination of interactions at the PP level and transcription factor activity at the PD level. This is supported by the observation that interactions in this cluster have an average shortest path of length 2 in the combined network but are more than three hops apart in the individual PD and PP networks.

Clusters G1 and G2 from the class of genetic interactions confirm our previous observation that these interactions consist of well-connected proteins in the PP network. This is apparent from the selection of the shortest-path STM across the PP and PP/PD networks and supported by the low average shortest-path distance. Additionally, both clusters also select long-distance paths

across the PD network, indicative of the lack of a common regulator.

Taken together, this indicates that each of the three functional interaction classes has a few rather distinct implementations of the interactions in the physical interaction network.

4 CONCLUSIONS AND OUTLOOK

We have shown that generalized STMs can be obtained based on diffusion kernels. Using these STMs, graph topological scale spaces can be constructed that characterize graph topology across the full range of scales, from local to a global scale. The measures described here already constitute a fairly complete set, covering centrality, clustering, paths, association and betweenness. Based on the principles outlined here, however, it should be straightforward to transform more existing topological measures into scale-aware versions.

Scale is an important aspect in relating physical networks to functional annotations and interactions. In a supervised classification setting, STMs improved performance in all our experimental settings, also in comparison with the equivalent standard (non-scale-aware) measures. To obtain good classification, local topology is clearly an important factor. However, using STMs, we also find that descriptors of higher scale were selected by the training algorithms.

We found that the different functional interaction classes each have their own topological signatures. These signatures are selected with remarkable robustness. Through unsupervised analysis, we showed that these classes of functional interactions are based on more fine-grained topological signals that are variations on the overarching topological signatures. Taken together, graph topological scale spaces clearly capture biologically meaningful features by exploiting, across multiple scales, graph topology of physical interaction data.

Application of STMs is not limited to protein networks. On the contrary, they have the potential to play an important role in answering various questions in the field of network biology. For instance, several studies have shown that topology-describing measures can be used to prioritize candidate disease genes (Ghersi *et al*., 2013; Köhler *et al*., 2008). There is also increasing evidence that the topological structure of brain connectivity determines many of its functions (Park and Friston, 2013). We believe that STMs and the associated graph topological scale spaces can provide new insights in many of these applications.

5 METHODS

5.1 Measures

Note that, below, we have omitted α for notational convenience and without loss of generality.

5.1.1 Shortest path The shortest-path measure s(x,y), i.e. the minimum number of vertices connecting node x and y, can be redefined to work across the scale-space by noting that K^β can be rewritten as

$$K^\beta = I + \beta H + \frac{\beta^2}{2!}H^2 + \frac{\beta^3}{3!}H^3 + \ldots \quad (2)$$

In this formulation, the contributions of different path lengths are represented as the different factors (H^i), with $(H^i)_{x,y} = 0$ for all $i < s(x,y)$.

As a result, when $\beta \to 0$, the diffusion signal approaches

$$K^\beta_{x,y} \sim \frac{r_{x,y}}{s(x,y)!}\beta^{s(x,y)} \ (\beta \to 0) \quad (3)$$

where $r_{x,y}$ is the number of shortest paths between x and y. All longer paths [i.e. $i > s(x,y)$] have negligible contributions, as $\beta^{s(x,y)} >> \beta^i$ because $\beta \to 0$.

Therefore, taking $-\log(K^\beta_{x,y})$ gives us $-\log(r_{x,y}) + \log(s(x,y)!) - s(x,y)\log(\beta)$. Again assuming that $\beta \to 0$, then $-\log(r_{x,y}) + \log(s(x,y)!)$ will become negligible, and we are left with $-s(x,y)\log(\beta)$. As $\log(\beta)$ is equal for all paths and negative as $\beta \to 0$, we remain with $s(x,y)$, which is the shortest-path length. For interpretation purposes, the $-log(\beta)$ factor can be divided out of the measure using the following scaling factor: $max(log\left(\frac{1}{\beta}\right), 1)$.

For higher values of β (i.e. an increase in scale), the shortest-path measure gradually transforms: through component $r_{x,y}$, the number of shortest paths between x and y will be taken into account. Moreover, because of the other components [with $i > s(x,y)$], longer paths will also gradually be incorporated. Intuitively, one may regard this as bundling (near) shortest paths, where more and more physical paths are covered as β increases. For values of β above 1.0, the focus gradually shifts from shortest paths to paths of certain length (owing to the balance between β^i and $i!$). Finally, for $\beta \to \infty$, and under the condition that there is a shortest path between x and y, $s^\beta_{x,y} \to \log(N_c(x))$, where $N_c(x)$ represents the number of nodes connected to x (and thus to y).

Shortest-path STM: $s^\beta(x,y) = -\log(K^\beta_{x,y})$

5.1.2 Jaccard index The Jaccard index is one of many available association measures (Fuxman Bass *et al*., 2013), describing how nodes are associated in terms of their shared set of interaction partners. It is defined as $J = \frac{n(x) \cap n(y)}{n(x) \cup n(y)}$, where n(x) and n(y) represent the set of neighbors of x and y, respectively.

In graph topological scale-space, all connected nodes become neighbors, so this definition cannot be used directly. Instead, we take into account the strength of each interaction. The intersection is subsequently defined as the minimum weight of both interactions per interaction partner ($\sum_i \min(K^\beta_{x,i}, K^\beta_{i,y})$) and the union as the maximum weight ($\sum_i \max(K^\beta_{x,i}, K^\beta_{i,y})$). This definition is equal to the original Jaccard for graphs in which all edges are equally weighted, yet with the definite advantage of also being applicable in graph topological scale spaces. For higher values of β, it will represent a Jaccard measure with increasingly larger neighborhoods, approaching 1 for $\beta \to \infty$.

Jaccard STM: $J^\beta(x,y) = \frac{\sum_i \min(K^\beta_{x,i}, K^\beta_{i,y})}{\sum_i \max(K^\beta_{x,i}, K^\beta_{i,y})}$

5.1.3 Degree and closeness centrality The degree centrality reflects the connectivity of a node in terms of the number of edges connected to it: deg(x). Again, using $\beta \to 0$, we find that $K^\beta_{x,x} \sim 1 - deg(x)\beta$. Consequently, $\frac{1 - K^\beta_{x,x}}{\beta}$ gives us a measure of the degree of each node. Increasing β gives us the degree with respect to increasing neighborhoods.

Closeness centrality is classically defined as $c(x) = \frac{1}{\sum_{i \setminus x} s(x,i)}$. It reflects the farness of a node x, by summing the shortest-path distances to all other nodes. This definition is inadequate for graphs with disconnected components. Therefore, an alternative formulation has been suggested (Dangalchev, 2006): $c(x) = \sum_{i \setminus x} 2^{-s(x,i)}$. Filling in the shortest-path STM, we obtain $c^\beta(x) = \sum_{i \setminus x} 2^{log(K^\beta_{x,i})} \sim \sum_{i \setminus x} K^\beta_{x,i} = 1 - K^\beta_{x,x}$. Apart from a scaling factor β, this is the same formulation as was obtained for the degree centrality STM. We will refer to this combined measure as the 'centrality'.

Centrality STM: $c^\beta(x) = \frac{1 - K^\beta_{x,x}}{\beta}$.

5.1.4 Clustering coefficient The clustering coefficient for a node is defined as the number of edges between its direct neighbors including itself, divided by the maximum number of possible edges, i.e. $cc(x) = \frac{2|e_x|}{deg(x)(deg(x)-1)}$, where $|e_x|$ is the number of edges among the direct neighbors of node x.

As was observed in the derivation of the Jaccard STM, in the graph topological scale-space the graph is fully connected. We propose to define a generalized scale-aware version of the clustering coefficient as the weighted average of Jaccard STM of all the interaction partners of x, i.e. $cc_\beta(x) = \sum_{i \setminus x} k(\beta)(x, i) * J_\beta(x, i)$. This formulation is not fully equivalent to the standard clustering coefficient when applied to a standard adjacency graph. A notable difference with the traditional measure is that edges are downweighted when a 'neighbor node' has many links to nodes outside the cluster. This conforms more closely to the common notion of an optimal cluster—high connectivity (low distance) between nodes within the cluster and low connectivity (large distances) to nodes outside the cluster. In the formulation proposed here, the size of the neighborhood (cluster) is defined by the scale parameter β.

Clustering coefficient STM: $cc^\beta(x) = \sum_{i \setminus x} K^\beta_{x,i} J^\beta(x, i)$

5.1.5 Betweenness centrality The betweenness centrality is defined as the number of shortest paths that pass through a node, i.e. $b(x) = \sum_{i,j \setminus x} \frac{q_{ij}(x)}{q_{ij}}$, where q_{ij} is the number of shortest paths between nodes i and j, and $q_{ij}(x)$ the number of those paths that pass through x.

In graph topological scale-space, the notion of a shortest path changes, as the graph is fully connected. Note, however, that if $s(x,y) = s(x,z) + s(z,y)$, we can infer that z is on the shortest path between x and y, by using the triangle inequality. As a result, we can exploit the previously defined shortest-path STM. Moreover, instead of discrete counting, we calculate a continuous score $s^\beta(x, y) - (s^\beta(x, z) + s^\beta(z, y))$ for each node pair x, y and normalize by the total number of pairs N^2, where N is the total number of nodes.

Betweenness centrality STM:
$b^\beta(z) = \frac{1}{N^2} \sum_{x,y} \left(s^\beta(x, y) - (s^\beta(x, z) + s^\beta(z, y))\right)$

5.1.6 From node to link-based STMs The STMs betweenness centrality, clustering coefficient and centrality are node-based. To extend their use to link-based classification and clustering, the average and the difference between the node-based STM values are used as measure. This results in eight link-based STMs (Jaccard and shortest-path STMs and the average and difference of the three node-based STMs).

5.1.7 From link to node-based STMs To perform node-based classification, link-based STMs were converted to node-based STMs. For the Jaccard measure, this was already accomplished through the clustering coefficient STM. Therefore, only the shortest-path measure was converted. This was done in a similar manner as for the clustering coefficient STM, i.e., $s^\beta(x) = \sum_{i \setminus x} K^\beta_{x,i} s^\beta(x, i)$

5.1.8 Values for β As values for β, we empirically choose a grid of size 20. Grid points were set according to $\frac{2^{8r}-1}{2^8-1} * (10.0 - 0.0001) + 0.0001$, with r = 0.0 ... 1.0 in 20 steps. This results in the following values for β: (0.0001, 0.0134, 0.0312, 0.0550, 0.0869, 0.1296, 0.1868, 0.2634, 0.3659, 0.5031, 0.6869, 0.9330, 1.2624, 1.7035, 2.2941, 3.0848, 4.1435, 5.5610, 7.4589, 10.0000)

5.2 Dimension reduction and clustering

The t-distributed stochastic neighbor embedding (t-SNE) dimension reduction algorithm was used with a perplexity value of 30 to reduce the 4M × 960 dataset to two dimensions. The t-transformed distances, which are used within the t-SNE algorithm to obtain a visualization, were clustered using k-means-based spectral clustering. We used $k = 50$. This was followed by a second round of hierarchical clustering, using single linkage, euclidean distance and a distance threshold of 1% of the data range.

5.3 Supervised analysis

The RNNC from the PRTools toolbox (Duin *et al.*, 2004) was used. This is a feed-forward neural network with an input layer that scales the data to unit variance, a layer of 100 sigmoid neurons performing a random rotation and shift, and an output layer. Forward feature selection was used to determine the optimal set of features. Classifier performance was determined using double-loop cross-validation and reported as the AUC. The inner loop was used for feature selection, whereas the outer loop was used to determine the final performance.

5.3.1 Protein function classification MIPS functional annotations were used to label proteins. To ensure the presence of enough positive instances in both testing and training sets, functional categories with <20 genes were not considered. In total, 16 MIPS functional classes (Fig. 2A) were considered for classification.

Classifier performance was determined for each pair of biological functions, using double-loop cross-validation, with a 5-fold inner and 4-fold outer loop. For each iteration of the outer loop, three folds were used for training/inner cross-validation, and the remaining fold was used for testing. This resulted in four AUC performance values (which were averaged) and their SD. Subsequently, these AUC and SDs were averaged for each functional category.

5.3.2 Interaction classification Gene pairs were labeled according to their functional interaction—no functional relation (neg), co-expression (coe), genetic (gen) or perturbation (pt). With these labels, six classification problems were investigated—neg-coe, neg-gen, neg-pt, coe-gen, coe-pt, gen-pt. To ensure a balanced classification experiment, the number of examples selected for the neg class was kept equal to the other class.

Classifier examples, represented by gene pairs, were divided into five folds. Because of the abundance of interactions available for training a classifier, we used only one fold for training and four folds for testing in each iteration. Feature selection was performed on one half of the training fold, and validation of these features was performed on the other half. This setup ensures that the data used for training and feature selection are completely non-overlapping between the iterations of the outer loop.

5.4 Feature selection

The total number of STMs (eight for link-based classification) combined with the 6 different values for α and 20 different values for β results in 960 features (topological descriptors). Forward feature selection was used to select the most informative features for specific classification tasks. The AUC score was used as performance measure. Briefly, using the double-loop cross-validation setup described before, we used the inner cross-validation loop to determine an optimal set of features. We thus obtained a list of selected features for each iteration of the outer cross-validation loop, which were then tested on the associated outer validation set (test set). Obtained performances were averaged. The different feature lists are summarized in Figure 3.

ACKNOWLEDGEMENT

The authors thank Laurens van der Maaten for a tailored implementation of the t-SNE algorithm.

Funding: The Netherlands Organisation for Scientific Research (NWO-Veni: 639.021.233).

Conflict of Interest: none declared.

REFERENCES

Barabási,A.-L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

Bonacich,P. (1972) Factoring and weighting approaches to status scores and clique identification. *J. Math. Sociol.*, **2**, 113–120.

Chua,G. *et al.* (2006) Identifying transcription factor functions and targets by phenotypic activation. *Proc. Natl Acad. Sci. USA*, **103**, 12045.

Costanzo,M. *et al.* (2010) The genetic landscape of a cell. *Science*, **327**, 425.

Dangalchev,C. (2006) Residual closeness in networks. *Physica A*, **365**, 556–564.

Duin,R. *et al.* (2004) Pr-tools. *Pattern Recognit. Tools*. http://www.prtools.org.

Estrada,E. (2010) Generalized walks-based centrality measures for complex biological networks. *J. Theor. Biol.*, **263**, 556–565.

Estrada,E. and Rodríguez-Velázquez,J.A. (2005) Subgraph centrality in complex networks. *Phys. Rev. E*, **71**, 056103.

Freeman,L.C. (1977) A set of measures of centrality based on betweenness. *Sociometry*, **40**, 35–41.

Fuxman Bass,J.I. *et al.* (2013) Using networks to measure similarity between genes: association index selection. *Nat. Methods*, **10**, 1169–1176.

Gasch,A.P. *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

Ghersi,D. *et al.* (2013) Disentangling function from topology to infer the network properties of disease genes. *BMC Syst. Biol.*, **7**, 1–12.

He,X. and Zhang,J. (2006) Why do hubs tend to be essential in protein networks? *PLoS Genet.*, **2**, e88.

Hu,Z. *et al.* (2007) Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet.*, **39**, 683–687.

Hughes,T. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Jansen,R. *et al.* (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res.*, **12**, 37–46.

Jordán,F. and Scheuring,I. (2002) Searching for keystones in ecological networks. *Oikos*, **99**, 607–612.

Joy,M. *et al.* (2005) High-betweenness proteins in the yeast protein interaction network. *J. Biomed. Biotechnol.*, **2005**, 96–103.

Katz,L. (1953) A new status index derived from sociometric analysis. *Psychometrika*, **18**, 39–43.

Kelley,R. and Ideker,T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.*, **23**, 561–566.

Kitano,H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662.

Köhler,S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.

Kondor,R.I. and Lafferty,J. (2002) Diffusion kernels on graphs and other discrete input spaces. In: *ICML* Vol. 2, pp. 315–322.

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

Mewes,H.-W. *et al.* (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, **32** (**Suppl. 1**), D41–D44.

Milenkoviæ,T. and Pržulj,N. (2008) Uncovering biological network function via graphlet degree signatures. *Cancer Inform.*, **6**, 257.

Ourfali,O. *et al.* (2007) Spine: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics*, **23**, i359–i366.

Park,H.-J. and Friston,K. (2013) Structural and functional brain networks: from connections to cognition. *Science*, **342**, 1238411.

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

Phillips,P.C. (2008) Epistasisthe essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.*, **9**, 855–867.

Ptacek,J. *et al.* (2005) Global analysis of protein phosphorylation in yeast. *Nature*, **438**, 679–684.

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

Stark,C. *et al.* (2010) PhosphoGRID: a database of experimentally verified *in vivo* protein phosphorylation sites from the budding yeast *Saccharomyces cerevisiae*. *Database*, **2010**, bau026.

Stark,C. *et al.* (2011) The biogrid interaction database: 2011 update. *Nucleic Acids Res.*, **39** (**Suppl. 1**), D698.

Teixeira,M. *et al.* (2006) The yeastract database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **34** (**Suppl, 1**), D446–D451.

Van der Maaten,L. and Hinton,G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 85.

van Wageningen,S. *et al.* (2010) Functional overlap and regulatory links shape genetic interactions between signaling pathways. *Cell*, **143**, 991–1004.

Westerhoff,H. and Palsson,B. (2004) The evolution of molecular biology into systems biology. *Nat. Biotechnol.*, **22**, 1249–1252.

Winterbach,W. *et al.* (2013a) Local topological signatures for network-based prediction of biological function. In: Ngom,A. *et al.* (eds) *Pattern Recognition in Bioinformatics*, Vol. 7986 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, pp. 23–34.

Winterbach,W. *et al.* (2013b) Topology of molecular interaction networks. *BMC Syst. Biol.*, **7**, 90.

Witkin,A.P. (1984) Scale-space filtering: A new approach to multi-scale description. In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84*. Vol. 9, IEEE, pp. 150–153.

Yu,H. *et al.* (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.*, **3**, e59.

Zhu,C. *et al.* (2009) High-resolution dna-binding specificity analysis of yeast transcription factors. *Genome Res.*, **19**, 556–566.

Zotenko,E. *et al.* (2008) Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput. Biol.*, **4**, e1000140.