JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2010) **7**, 423–437
doi:10.1098/rsif.2009.0192
Published online 22 July 2009

Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data

Tijana Milenković[1], Vesna Memišević[1], Anand K. Ganesan[2] and Nataša Pržulj[1,*]

[1]*Department of Computer Science, University of California, Irvine, CA 92697-3435, USA*
[2]*Department of Dermatology and Biological Chemistry, University of California, Irvine, CA 92697-2400, USA*

Many real-world phenomena have been described in terms of large networks. Networks have been invaluable models for the understanding of biological systems. Since proteins carry out most biological processes, we focus on analysing protein–protein interaction (PPI) networks. Proteins interact to perform a function. Thus, PPI networks reflect the interconnected nature of biological processes and analysing their structural properties could provide insights into biological function and disease. We have already demonstrated, by using a sensitive graph theoretic method for comparing topologies of node neighbourhoods called 'graphlet degree signatures', that proteins with similar surroundings in PPI networks tend to perform the same functions. Here, we explore whether the involvement of genes in cancer suggests the similarity of their topological 'signatures' as well. By applying a series of clustering methods to proteins' topological signature similarities, we demonstrate that the obtained clusters are significantly enriched with cancer genes. We apply this methodology to identify novel cancer gene candidates, validating 80 per cent of our predictions in the literature. We also validate predictions biologically by identifying cancer-related negative regulators of melanogenesis identified in our siRNA screen. This is encouraging, since we have done this solely from PPI network topology. We provide clear evidence that PPI network structure around cancer genes is different from the structure around non-cancer genes. Understanding the underlying principles of this phenomenon is an open question, with a potential for increasing our understanding of complex diseases.

Keywords: biological networks; protein interaction networks; network topology; cancer gene identification

1. INTRODUCTION

1.1. Background

Large amounts of biological network data have become available owing to recent advances in experimental biology. Networks are invaluable models for better understanding of biological systems (Barabási & Oltvai 2004). To understand living cells, one needs to study them as interconnected systems rather than as a collection of individual parts (Ideker & Sharan 2008). Whether the constituents of a network are molecules, cells or living organisms, the network provides a framework to model the complex events that emerge from interactions among these parts.

*Author for correspondence (natasha@ics.uci.edu).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2009.0192 or via http://rsif.royalsocietypublishing.org.

Nodes in biological networks represent biomolecules such as genes, proteins or metabolites, and edges connecting these nodes indicate functional, physical or chemical interactions between the corresponding biomolecules. Understanding these complex biological systems has become an important problem that has led to intensive research in network analyses, modelling, and function and disease gene identification and prediction. The hope is that using such systems-level approaches to analysing and modelling complex biological systems will provide insights into the inner working of the cell, biological function and disease.

Because it is the proteins that execute the genetic code and carry out most biological processes, we focus on protein–protein interaction (PPI) networks. In these networks, nodes correspond to proteins and undirected edges represent physical interactions among them. As proteins are essential macromolecules of life,

Received 19 May 2009
Accepted 24 June 2009

understanding their function and role in disease is of great importance.

Methods for protein function detection have shifted their focus from targeting individual proteins based solely on sequence homology to systems-level analyses of entire proteomes based on PPI network topology (Sharan *et al.* 2007; Milenković & Pržulj 2008). Since proteins interact to perform a certain function rather than functioning in isolation, these networks of protein interactions by definition reflect the interconnected nature of biological processes. Therefore, analysing structural properties of PPI networks may provide useful clues about the biological function of individual proteins, protein complexes, pathways they participate in and larger subcellular machines. For example, proteins that are closer in a network are more likely to perform the same function (Sharan *et al.* 2007). In the most simple form, this assumption has been used to investigate the direct neighbourhood of an unannotated protein, and annotate it with the most common functions among its annotated neighbours (Schwikowski & Fields 2000). Other examples include more recent studies demonstrating that proteins with similar topological neighbourhoods have similar biological characteristics (Guerrero *et al.* 2008; Milenković & Pržulj 2008).

Similarly, owing to the increase in availability of human protein interaction data, the focus of bioinformatics in general has shifted from understanding the networks of model species, such as yeast, to understanding the networks responsible for human disease (Ideker & Sharan 2008). These studies have been trying to address several challenges: investigating network properties of disease genes, identifying gene–disease or drug–drug target associations and predicting novel disease genes.

There is an open debate as to whether genes involved in serious diseases such as cancer can be distinguished based on their properties and position in a PPI network. For example, cancer genes have been shown to have greater connectivities and centralities compared with non-cancer genes (Jonsson & Bates 2006). However, the relationship between disease genes and their network degrees might need a more careful consideration, as most of the disease genes do not show a tendency to code for proteins that are hubs (Goh *et al.* 2007).

Radivojac *et al.* (2008) have tried to identify gene–disease associations by encoding each gene in a PPI network based on the distribution of shortest path lengths to all genes associated with disease or having known functional annotation. Moreover, PPI networks have recently been combined with the networks describing the relationships between diseases and genes causing them (Goh *et al.* 2007), as well as between drugs and their protein targets (Yidirim *et al.* 2007), thus giving new insights into pharmacology.

Finally, studies trying to predict involvement of genes in diseases such as cancer have been relying on the key assumption that a neighbour of a disease-causing gene in a PPI network is likely to cause either the same or a similar disease (Ideker & Sharan 2008). For example, Aragues *et al.* (2008) started from the hypothesis that proteins whose partners have been annotated as cancer genes are likely to be cancer genes as well, constructed a cancer protein interaction network composed of known cancer genes and their direct interacting partners and demonstrated that the 'cancer linker degree' of a protein, i.e. the number of its cancer-related neighbours in this network, is a good indicator of the probability that the gene is a cancer gene.

1.2. Our study

Defining the relationship between PPI network topology and protein function and disease has been one of the major challenges in the post-genomic era. Here, we further explore this relationship, focusing on cancer in particular. We investigate if topological properties of PPI networks can be used to imply involvement of proteins in cancer. Unlike other approaches that have been relying on the assumption that network neighbours of cancer genes are also involved in cancer (Aragues *et al.* 2008), we test a different hypothesis: do the genes that are involved in cancer have similar 'topological signatures' (defined below) without necessarily being adjacent in the network? Furthermore, other studies rely only on global network properties, such as high node degrees, to characterize cancer genes and they generalize this to the entire set of cancer genes in a network (Jonsson & Bates 2006). In contrast, we rely on a highly constraining local network measure that describes network interconnectedness of up to '4-deep neighbourhood' of a node (see below). Furthermore, we do not assume that all cancer genes should have similar topological signatures. Instead, we allow for a possibility that proteins involved in different cancers might have different network neighbourhoods.

We have already used a sensitive graph theoretic method for comparing local structures of node neighbourhoods to demonstrate that in PPI networks, biological function of a node and its local network structure are closely related (Milenković & Pržulj 2008). The method summarizes a protein's local topology in a PPI network into its 'signature'. Then, signature similarities between all protein pairs are computed, measuring topological resemblance of their neighbourhoods. It has been shown that clusters obtained by grouping topologically similar proteins under the signature similarity measure are statistically significantly enriched in biological function, membership in protein complexes, subcellular localization and tissue expression (Milenković & Pržulj 2008; Pržulj & Milenković in press). Owing to similarity in biochemical manifestations of different diseases and various types of cancer in particular, here we apply our approach to explore whether cancer genes share similar topological signatures as well. More specifically, we apply a series of clustering methods to proteins based on their topological signature similarities and analyse whether the obtained clusters are statistically significantly enriched with cancer genes. Thus, the novelty of our approach is the evaluation of different clustering algorithms and application of our method to cancer. Based on this approach, we predict novel cancer gene candidates,

validating about 80 per cent of our predictions in the literature.

Furthermore, we provide biological application and validation of our predictions. RNAi-based functional genomics is an unbiased approach to identify genes that specifically regulate cellular phenotypes (Whitehurst *et al.* 2007; Krishnan *et al.* 2008; Silva *et al.* 2008). We have previously used this approach to identify novel regulators of melanogenesis in human cells, a differentiated cellular phenotype (Ganesan *et al.* 2008). Previous studies have postulated that oncogenes negatively regulate melanin production and cellular differentiation (Halaban 2002). We use our network topology-based approach to identify negative regulators of melanogenesis identified in our siRNA screen that are also cancer genes. Twenty-seven putative cancer genes were identified in this dataset, 85 per cent of which are linked to cancer through literature search. Among these genes are known negative regulators of melanogenesis, further demonstrating the power of network topological signatures to specifically identify cancer genes in biologically relevant datasets.

We compare the performance of our method with that of Aragues *et al.* (2008), which also predicts from PPI networks the involvement of genes in cancer. While Aragues *et al.* (2008) focus only on direct network neighbours of cancer genes, we account for complex wirings of their up to 4-deep neighbourhoods, as depicted in figure 1*a*; we demonstrate that out of all known cancer gene pairs that have similar topological signatures, 96 per cent are *not* direct neighbours in the PPI network. Moreover, in addition to network topology, Aragues *et al.* (2008) also use gene expression data and structural and functional properties of cancer proteins, while we use the network topology only. Even though we do not use any information external to PPI network topology, our approach is superior, as it results in higher prediction accuracy. Thus, graphlet degree signatures provide a better prediction accuracy than less constraining network properties such as nodes' direct neighbours, even when nodes' direct neighbourhoods are integrated with other data types.

2. APPROACH

We use 'graphlet degree signatures' (Milenković & Pržulj 2008) of proteins in a PPI network to predict their involvement in cancer. *Graphlets* are small connected induced subgraphs of a large network (figure 1*a*; Pržulj *et al.* 2004*a*). The method generalizes the degree of a node, which counts how many edges the node touches, into the 'vector of graphlet degrees', or the 'node signature' (or just 'signature' for brevity), which counts how many graphlets of a given type, such as a triangle or a square, the node touches (figure 1*b*). All two- to five-node graphlets, presented in figure 1*a*, are taken into account. Thus, the signature of a node describes the topology of its neighbourhood and captures the node's interconnectivities with its up to four-neighbours. Next, 'signature similarities' are computed for each pair of proteins in a PPI network (see §3); higher signature similarity corresponds to higher topological similarity of neighbourhoods of two nodes (Milenković & Pržulj 2008).

To increase the coverage of PPIs, the human PPI network that we analyse is the union of the human PPI networks from HPRD (Peri *et al.* 2003), BIOGRID (Stark *et al.* 2006) and Radivojac *et al.* (2008), consisting of 47 303 physical interactions among 10 282 proteins. When protein signatures are computed, all proteins are taken into consideration. However, in all of our subsequent analyses, we focus only on proteins with more than three interacting partners, because poorly connected proteins are more likely to be involved in noisy interactions. Similar was done by Brun *et al.* (2004) and Milenković & Pržulj (2008). In the human PPI network, there are 5423 proteins with degrees higher than 3.

Using signature similarities as the distance measure, we cluster proteins in the human PPI network and analyse the enrichment of cancer genes in these clusters. If the cluster containing a gene that is currently not reported to be involved in cancer contains many known cancer genes, it is likely that the gene is also involved in cancer. We denote as 'known cancer genes' the set of genes implicated in cancer that is available from the following databases: Cancer Gene Database,[1] Cancer Genome Project—the Cancer Gene Census (Futreal *et al.* 2004),[2] GeneCards (Safran *et al.* 2002),[3] Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa & Goto 2000)[4] and Online Mendelian Inheritance in Man (OMIM; Hamosh *et al.* 2002).[5] Cancer Gene Database contains a list of genes involved in diseases derived from Medline abstracts by mixture of automatic text mining, semi-automatic verification and manual validation/scoring of results; out of all disease genes, we extract those described as being involved in cancer. Cancer Gene Census contains those genes from the literature for which mutations have been causally implicated in cancer. GeneCards provides the list of genes that are related to cancer in any of the following databases: SWISS-PROT, Genatlas, GeneTests, GAD, GDPInfo, bioalma, Leiden, Atlas, BCGD, TGDB and HGMD. Out of all disease genes in OMIM, we extract those described as being involved in cancer by at least two studies. Cancer genes from KEGG are those that are members of known cancer pathways originating from the literature. There are 1688 unique known cancer genes in the merged dataset, out of which 1205 are found in the PPI network and 679 out of these 1205 genes have degrees higher than 3, representing our final list of 'known cancer genes'.

There is no clustering algorithm that can be universally used to solve all problems. Various clustering algorithms have been proposed, originating from different research communities and aiming to solve different problems, each with its own advantages and disadvantages (Xu & Wunsch 2005). Therefore, we test the

[1]http://ncicb.nci.nih.gov/projects/cgdcp.
[2]http://www.sanger.ac.uk/genetics/CGP/Census/.
[3]http://www.genecards.org/.
[4]http://www.genome.jp/kegg/disease/.
[5]http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim.

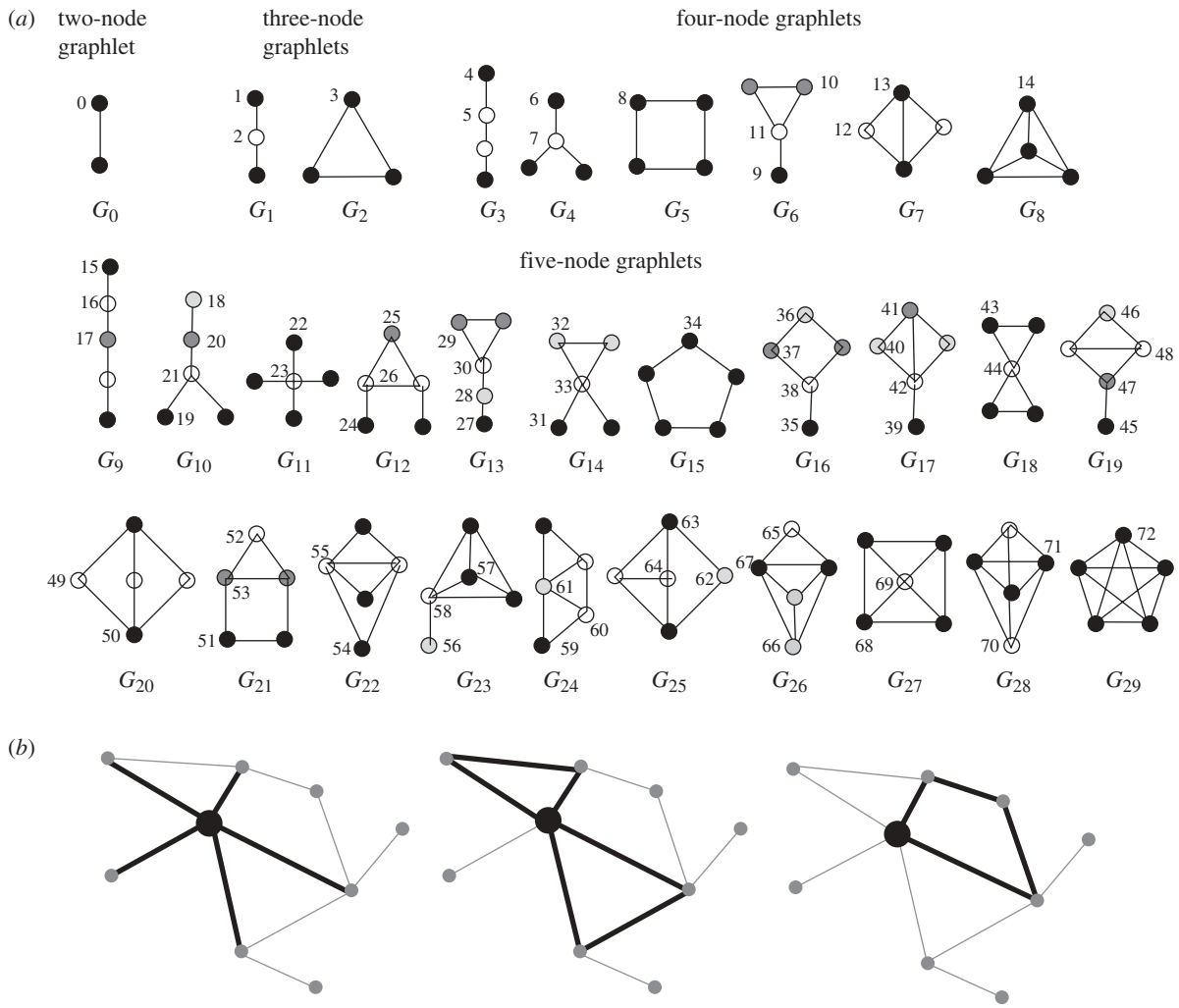

Figure 1. (*a*) Automorphism orbits 0, 1, 2, ..., 72 for the 30 two-, three-, four- and five-node graphlets G_0, $G_1, \ldots, G_{29}$. In a graphlet G_i, $i \in 0, 1, \ldots, 29$, nodes belonging to the same orbit are of the same shade. Adapted from Pržulj (2007). (*b*) An illustration of how the degree of the large black node in the leftmost panel is generalized into its signature that counts the number of different graphlets that the node touches, such as triangles (the middle panel) or squares (the rightmost panel).

performance of different clustering algorithms applied to our signature similarity measure. We evaluate four different clustering methods: hierarchical (HIE), K-medoids (KM), K-nearest neighbours (KNN) and signature threshold-based clustering (ST). Each method differs in the way clusters are formed (see §3). For example, whereas KNN and ST allow for overlap between clusters, HIE and KM do not. HIE and KM methods require the number of clusters to be prespecified, KNN requires the size of clusters as the input parameter and ST depends on the choice of the signature similarity threshold (SST). For each of the clustering methods, we use different parameters to test how the accuracy of the method changes with the parameter choice (see §3).

For each of the four clustering methods and their corresponding parameters, we do the following. After clustering the network, for each protein, we compute the enrichment of known cancer genes in the cluster the protein belongs to and assess the statistical significance of observing the given enrichment (see §3). We discard the cluster from further analysis if the probability of observing the enrichment purely by chance is higher than 0.08. Otherwise, we predict whether the protein for which the cluster was formed is involved in cancer or not. We define a protein to be a cancer gene if the known cancer gene enrichment in its cluster is both statistically significant and above a given enrichment threshold (also called hit-rate threshold, or HRT; see §3 for details). We vary HRT from 10 to 90 per cent, in increments of 10 per cent. We repeat the above procedure for each protein in the network.

For each clustering method, each corresponding parameter and each HRT, we evaluate the prediction accuracy of the method by using leave-one-out cross-validation and standard measures of precision and recall (see §3). The leave-one-out cross-validation that hides the knowledge about a single protein at a time and predicts it using the PPI network and the knowledge about all other proteins is commonly used to evaluate the prediction accuracy of methods for cancer gene prediction (Aragues *et al.* 2008) or protein function prediction (Sharan *et al.* 2007). Intuitively, precision can be seen as a measure of exactness of a prediction method, whereas recall is a measure of

completeness of the method. To simplify the comparison of different clustering methods, precision and recall are combined into a commonly used *F*-score (see §3). We compute *F*-scores for all clustering methods, their corresponding parameters and HRTs. We report *F*-scores only for the best parameter choice for a given method, across all HRTs.

To assess the significance of observing given *F*-scores, we compare *F*-scores obtained when predictions are made from real data with *F*-scores obtained when predictions are made from randomized data (see §3). Thus, if protein signatures and signature similarities indeed capture true biological signal, it is expected that *F*-scores for randomized data will be lower than those for real data.

We provide the resulting list of cancer gene predictions (see §4). To further demonstrate the correctness and validity of these predictions, we perform a literature search and identify studies that have linked our predictions to cancer. We also validate our predictions biologically by finding among negative regulators of melanogenesis identified in our siRNA screen those genes that are also involved in cancer. Finally, we compare our results with those of related studies to demonstrate the superiority of our approach.

3. METHODS

3.1. Graphlet degree signatures and signature similarities

To predict the involvement of genes in cancer, we apply the similarity measure of nodes' local neighbourhoods, as described by Milenković & Pržulj (2008). This measure of node topological similarity generalizes the degree of a node, which counts the number of edges that the node touches, into the vector of *graphlet degrees*, counting the number of graphlets that the node touches; *graphlets* are small connected non-isomorphic induced subgraphs of a large network (Pržulj *et al.* 2004*a*). The method counts the number of graphlets touching a node for all two- to five-node graphlets, denoted by G_0, $G_1, \ldots, G_{29}$ in figure 1*a*. Clearly, the degree of a node is the first coordinate in this vector, since an edge (graphlet G_0) is the only two-node graphlet. This vector is called the *signature* of a node. To take into account the symmetry groups within a graphlet, the notion of *automorphism orbits* (or just *orbits*, for brevity) is used for all graphlets with two to five nodes. For example, it is topologically relevant to distinguish between nodes touching a three-node linear path (graphlet G_1) at an end or at the middle node. By taking into account these symmetries between nodes of a graphlet, there are 73 different orbits for two- to five-node graphlets, numerated from 0 to 72 in figure 1*a* (see Pržulj (2007) for details). Thus, the signature vector of a node, describing its up to 4-neighbourhood, has 73 coordinates.

The node signature similarities are computed as follows (Milenković & Pržulj 2008). For a node u, u_i denotes the ith coordinate of its signature vector, i.e. u_i is the number of times node u touches an orbit i. The distance $D_i(u, v)$ between the ith orbits of nodes u and v is defined as: $D_i(u, v) = w_i \times (|\log(u_i + 1) - \log(v_i + 1)|)/\log(\max\{u_i, v_i\} + 2)$, where w_i is a weight of orbit i signifying its 'importance' (see Milenković & Pržulj (2008) for details). The total distance $D(u, v)$ between nodes u and v is defined as: $D(u, v) = \sum_{i=0}^{72} D_i / \sum_{i=0}^{72} w_i$. The distance $D(u, v)$ is in $[0, 1)$, where distance 0 means that signatures of nodes u and v are identical. Finally, the *signature similarity*, $S(u, v)$, between nodes u and v is: $S(u, v) = 1 - D(u, v)$ (see Milenković & Pržulj (2008) for details).

Obviously, higher signature similarity corresponds to higher topological similarity of neighbourhoods of two nodes. In figure 2*a*, we illustrate neighbourhoods of two known cancer genes, *ZNF*384 and *DDX*6, that have a high signature similarity of 0.97; note that the shortest path distance in the PPI network between these proteins is 4. Additionally, in figure 2*b*, we present signatures of four known cancer genes, where pairs of genes (*ZNF*384, *DDX*6) and (*TP*53, *BRCA*1) have high signature similarities of 0.97 and 0.96, respectively, as clearly indicated by their very similar signature vectors in the figure; however, all remaining protein pairs in the figure {(ZNF384, TP53), (ZNF384, BRCA1), (DDX6, TP53), (DDX6, BRCA1)} have very low signature similarities of below 0.25, as indicated by their very different signature vectors.

3.2. Clustering methods

We cluster proteins based on their signature similarities. Intuitively, proteins with high signature similarities should be clustered together, whereas proteins with lower signature similarities should not. To calculate signature vectors of proteins, the entire PPI network with all interactions is taken into account. However, for clustering, poorly connected proteins with fewer than four interacting partners are discarded. Moreover, only clusters with three or more proteins are taken into consideration. Four different clustering methods have been used: hierarchical (HIE), K-medoids (KM), K-nearest neighbours (KNN) and signature threshold-based clustering (ST).

3.2.1. Hierarchical clustering (*HIE*). With this method, a cluster tree, or dendrogram, is created. The tree is not a single set of clusters, but a multi-level hierarchy, where clusters at one level are joined as clusters at the next level. Leaves of the tree are proteins in the PPI network and an interior node in the tree represents a cluster made up of all children of the node. The algorithm steps are as follows: (1) assign each protein to its own cluster, (2) find the 'closest' pair of clusters and merge them into a single cluster; in the initial step, the closest pair of clusters will be the pair of proteins with the highest signature similarity; in case there is more than one such pair, a pair is selected randomly from all of the closest pairs, (3) compute the 'closeness' between the newly formed cluster and each of the old clusters; the closeness between the new cluster and an old cluster is the average of signature similarities between proteins of the new cluster and proteins of the old cluster, and (4) repeat the

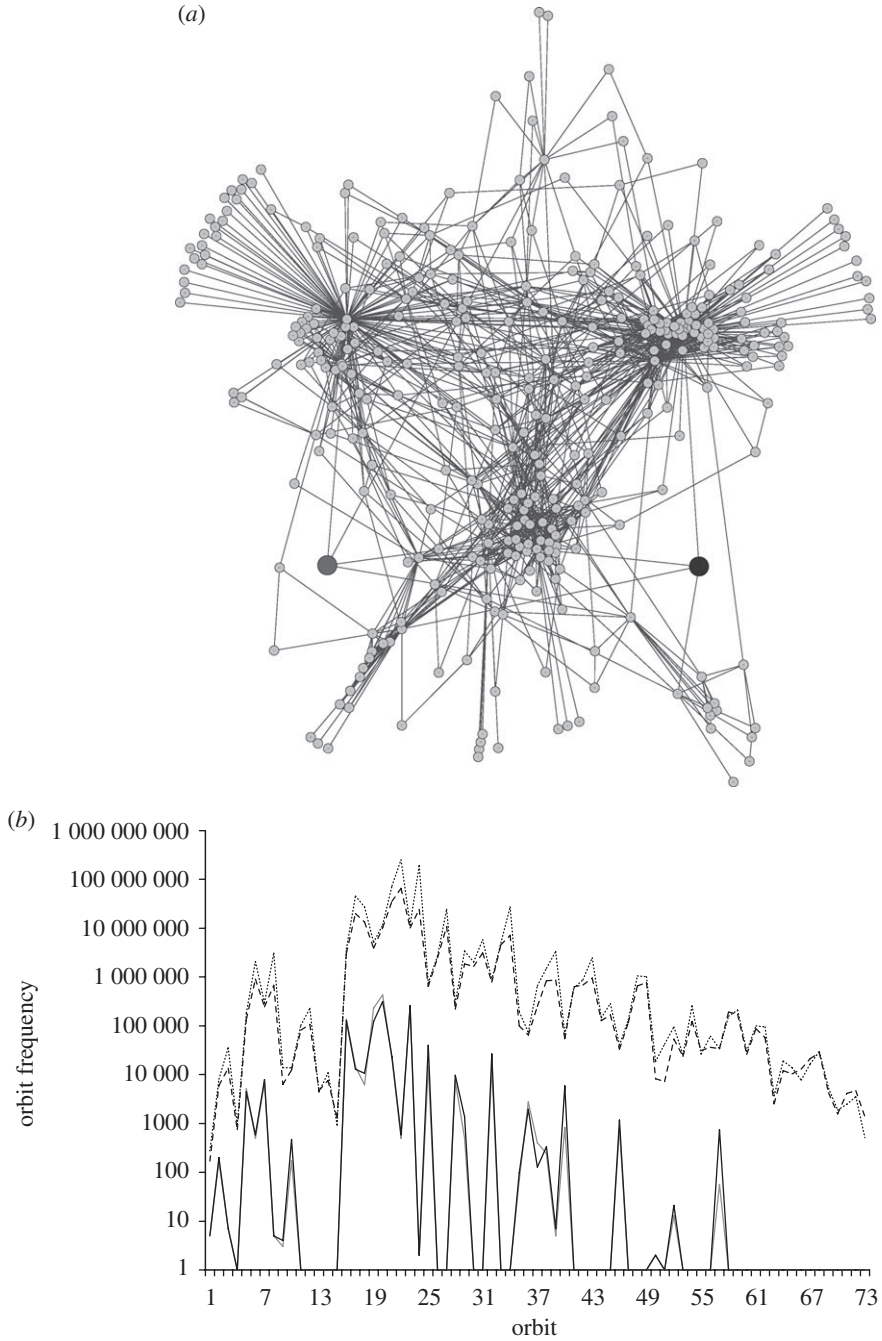

Figure 2. Illustration of node signatures and signature similarity measure. (*a*) 2-deep network neighbourhoods of proteins ZNF384 (large grey node) and DDX6 (large black node) that have a signature similarity of 0.97. (*b*) Signature vectors of protein pairs (ZNF384 (grey), DDX6 (black)) and (TP53 (dotted), BRCA1 (dashed)) with signature similarities above 0.95. The 73 orbits are presented on the abscissa and the numbers of times that nodes touch a particular orbit are presented on the ordinate in log scale. In the interest of the aesthetics of the plot, we added 1 to all orbit frequencies to avoid the log-function going to infinity in the case of orbit frequencies of 0.

two previous steps until all proteins are clustered into a single cluster. Hierarchical clustering does not explicitly require an *a priori* specified number of clusters. However, to perform any analysis, it is necessary to create partition of K_{H} disjoint clusters, cutting the hierarchical tree at some point. We cut the hierarchical tree at different points to produce a different number of clusters. We use the following values for K_{H}: 100, 250, 500, 750, 1000, 1250, 1500, 1750, 2000, 2250 and 2500.

3.2.2. K-medoids clustering (KM). KM is a modification of the classic K-means algorithm that chooses actual data points as *centres*, i.e. medoids; a *medoid* is the 'central' data point of a cluster whose average distance to all other data points in the cluster is minimal. The algorithm steps are as follows: (1) pick K_{KM} proteins as centres of K_{KM} clusters and assign all the remaining proteins in the PPI network to these centres; each protein will be assigned to the centre that has minimal signature 'distance' to it;

signature distance is equal to 1 − signature similarity; solve any ties randomly, (2) in each cluster C, pick protein X as the new centre of the cluster, so as to minimize the total sum of signature distances between protein X and all other proteins in cluster C, (3) reassign all proteins to new centres as explained in step (1), and (4) repeat the previous two steps until the algorithm converges, i.e. until the same set of centres is being chosen from one iteration to the next. We tried the following values for K_{KM}: 100, 250, 500, 750, 1000, 1250, 1500, 1750, 2000, 2250 and 2500. However, the algorithm could not converge for values of K_{KM} of 1500 or higher. Additionally, the set of clusters obtained with KM depends drastically on the choice of initial centres. Thus, for each K_{KM}, we repeated the algorithm 10 times, and we accepted the value of K_{KM} as valid only if the same set of clusters was obtained in at least 50 per cent of runs. All K_{KM}s of 750 or lower did not produce consistent sets of clusters in more than 50 per cent of runs. Thus, we further analysed only clusters obtained with K_{KM}s of 1000 and 1250; more specifically, we analysed the set of clusters that was obtained in more than 50 per cent of runs for a given K_{KM}.

3.2.3. K-nearest neighbours clustering (*KNN*). For each protein in the PPI network (with degree higher than 3), we create the cluster containing that protein and its $K_{\mathrm{KNN}} - 1$ closest neighbours, i.e. $K_{\mathrm{KNN}} - 1$ proteins that have the highest signature similarities with the protein; ties are broken randomly. Thus, for each protein, the resulting cluster will contain K_{KNN} proteins, including the protein of interest itself. We used the following K_{KNN} values: 3, 5, 8, 13, 21, 34, 55, 89, 144 and 233. Clearly, unlike HIE and KM, KNN allows for overlap between clusters.

3.2.4. Signature threshold-based clustering (*ST*). For each protein, we identify the cluster containing that protein and all other proteins in the network that have signature similarities with it above a certain threshold. We use the following SSTs: 0.7, 0.8, 0.85, 0.9, 0.925, 0.95, 0.975 and 1. Note that by approximating the distribution of signature similarities between all protein pairs in the PPI network with the normal distribution, and by finding Z-scores and their corresponding p-values for different SSTs, we find that the statistically significant SST is 0.85, with p-value of 0.045. For this reason, we do not find it necessary to analyse SSTs below 0.7.

3.3. Statistical significance and prediction accuracy

For a given clustering method and a given parameter, we measure the prediction accuracy of the method by using leave-one-out cross-validation: a single gene for which the prediction is made is used as the validation data, and all remaining genes based on which we make the prediction are used as the training data; we repeat the procedure for each gene. For each gene, we identify the cluster formed for that gene (for KNN and ST), or we identify the cluster containing that gene (for KNN and KM). We compute 'hit-rate' of each cluster, where hit-rate is defined as the percentage of cancer genes in the cluster out of all genes in the cluster, excluding the protein of interest. We predict a gene to be cancer-related if the hit-rate in its cluster is above a given threshold; in this process, we hide the information whether the gene of interest is cancer-related or not, and we do not count this gene towards cancer gene enrichment in its cluster. By repeating this procedure for each gene, and by varying the enrichment threshold, we measure how many of the genes were correctly classified. We do so by using standard precision and recall measures, as explained below.

We also compute the statistical significance (p-value) of observing a given hit-rate in the cluster, measuring the probability that the cluster was enriched by a given number of cancer genes purely by chance. This probability is computed as follows: the total number of proteins in the PPI network with degrees higher than 3 is 5423; the size of a cluster of interest, C, is $|C|$; the number of proteins in cluster C that are known cancer genes, excluding the protein of interest, is k; there are 679 known cancer genes with degrees higher than 3 in the entire PPI network. Then, the hit-rate of cluster C is $k/(|C| - 1)$, and the p-value for cluster C, i.e. the probability of observing the same or higher hit-rate purely by chance, is

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{679}{i}\binom{5423-679}{|C|-i}}{\binom{5423}{|C|}}.$$

Depending on a method and its application, sensible cut-offs for p-values were reported to range from 10^{-2} to 10^{-8} (King *et al.* 2004). We continue our analysis only for proteins for which cancer gene enrichment in their corresponding clusters is 0.08 or lower. For these proteins, we predict them as being cancer-related if the hit-rate in their corresponding clusters is above a given HRT. We vary HRTs from 10 per cent to 90 per cent in increments of 10 per cent. Thus, we make a different set of predictions for each clustering method, each parameter and each HRT.

Given a set of predictions, we measure the prediction accuracy of the method by using standard measures of precision and recall, combined into a commonly used F-score. Precision can be seen as a measure of exactness, whereas recall is a measure of completeness. Given that we produce n cancer gene predictions for a given method, parameter and HRT, *precision* is the number of known cancer genes that are in our n predictions divided by n. Thus, it measures the number of true positives (tp) out of both true positives (tp) and false positives (fp). Precision = tp/(tp + fp). *Recall*, on the other hand, is the number of known cancer genes in our n predictions divided by the total number of known cancer genes in the PPI network (with degrees higher than 3). Thus, it measures the number of true positives divided by the sum of the number of true positives and false negatives (fn). Recall = tp/(tp + fn). Note that we assume that currently unreported cancer genes are indeed non-related to cancer; this

assumption will certainly not hold as new cancer genes are being identified. Once precision and recall are computed for a given clustering method, parameter and HRT, *F*-score is computed as follows: $F\text{-score} = 2 \cdot \text{Precision} \cdot \text{Recall}/(\text{Precision} + \text{Recall})$. We use *F*-score since this combined measure of precision and recall makes it easier to evaluate different clustering methods against each other compared with using precision and recall measures individually.

Finally, to asses the significance of observing given *F*-scores, we compare *F*-scores obtained for real data with *F*-scores obtained when randomized data are used. By randomized data, we mean that we randomly shift labels of proteins in the PPI network before we compute signature similarities and perform any clustering; it is equivalent to say that we randomly permute signature similarities between all protein pairs in the network. Then, we repeat the whole procedure for clusters formed from this randomized data: we compute their hit-rates and *p*-values, make cancer gene predictions and measure *F*-scores.

4. RESULTS AND DISCUSSION

4.1. Results

Topological properties of PPI networks have already been linked to biology. For example, it has been proposed that the phenotypic consequence of a single gene deletion is correlated with the degree of its protein product in the PPI network (Jeong *et al.* 2001), although the subsequent studies questioned the observed correlation (He & Zhang 2006; Zotenko *et al.* 2008). If we accept the lethality–centrality hypothesis, then the topology around essential genes in a PPI network can be described by orbits 0, 2, 7, 21, 23 and 33 in figure 1*a*. Another example are protein complexes, which are believed to correspond to dense subgraphs in a PPI network (Pržulj *et al.* 2004*b*; Sharan and Ideker 2006; Sharan *et al.* 2007). Thus, the topology of protein complexes in PPI networks can be described by orbits 3, 12–14 and 65–72 in figure 1*a*. Biochemical pathways have also been linked to PPI network topology and are believed to correspond to sparse network regions (Pržulj *et al.* 2004*b*; Sharan and Ideker 2006). Thus, they can be described by orbits 1, 2, 4–7 and 15–23 in figure 1*a*. Note that the complex wiring of cellular networks implies the existence of an overlap between orbits involved in lethal proteins, protein complexes and pathways.

We further explore the link of PPI network topology and biology. We do not focus on any particular graphlet or orbit alone, since in isolation they can give insight into only a slice of biological information. Instead, we integrate all possible two- to five-node graphlets and all of their orbits into a highly constraining measure of network structural similarity between nodes, graphlet degree vectors. We show that such a detailed measure of network topology provides an insight into complex biological mechanisms, such as protein function and involvement in disease, that could not have been inferred from weaker measures of network topology, or biological information external to PPI network topology, such as protein sequence (Milenković & Pržulj 2008; Kuchaiev *et al.* 2009). We also demonstrate that this measure of network topology is capable of successfully identifying novel cancer gene candidates from PPI network topology alone.

Comparison of the prediction accuracy of the four clustering methods with respect to *F*-scores is presented in figure 3*a*. *F*-scores are shown only for the best parameter choice for each of the methods. These parameters are $K_{\text{H}} = 1250$, $K_{\text{KM}} = 1000$, $K_{\text{KNN}} = 8$ and $\text{SST} = 0.95$ for HIE, KM, KNN and ST, respectively (see §3). Other parameters that we tested for a given clustering method produced lower *F*-scores over the entire HRT range. Overall, KNN is the best out of the four clustering methods, followed by KM, ST and HIE, respectively.

The superiority of KNN over KM is not surprising. With KNN, a protein of interest is clustered with the top K most signature-similar proteins in the network. With KM, a protein of interest is clustered with a signature-closest centre protein, i.e. with the centre protein having the highest signature similarity with it (see §3). However, the signature-closest centre protein is not necessarily the most signature-similar protein in the network. Additionally, KNN allows for overlap between clusters, whereas KM does not. Thus, with KNN, known cancer genes can be positioned in multiple clusters, and therefore the number of possible clusters that are significantly enriched with cancer genes might be higher for KNN than for KM. Additionally, proteins perform the function or participate in a disease by interacting with other proteins within a functional module, but also with proteins across modules. Thus, it might be biologically relevant to allow for the overlap between clusters. The better performance of KNN over ST could be explained as follows. With ST, proteins with signature similarities above a given threshold are clustered with a protein of interest. However, a fixed threshold is used for all proteins in the network. A fixed threshold might be too stringent for proteins with complex and dense neighbourhoods, not allowing biologically relevant proteins to be clustered together just because their similarities are below the threshold. On the other hand, it might be too flexible for proteins with sparse neighbourhoods, allowing for too many potentially biologically unrelated proteins to be clustered together. Finally, the worst performance of HIE over all other methods could also be explained by its lack of overlap between clusters. Moreover, when the total signature similarity between two clusters is computed (see §3), the average of signature similarities between all pairs of proteins across the two clusters is used, thus potentially damaging the quality of clustering.

Each clustering method has its advantages and disadvantages. We compared the performance of different algorithms with respect to *F*-score. However, the choice of the most appropriate clustering strategy is far from being simple and is an important research problem (Xu & Wunsch 2005). Therefore, it is difficult to determine which of the four clustering methods is indeed the best for our application. Thus, we produce

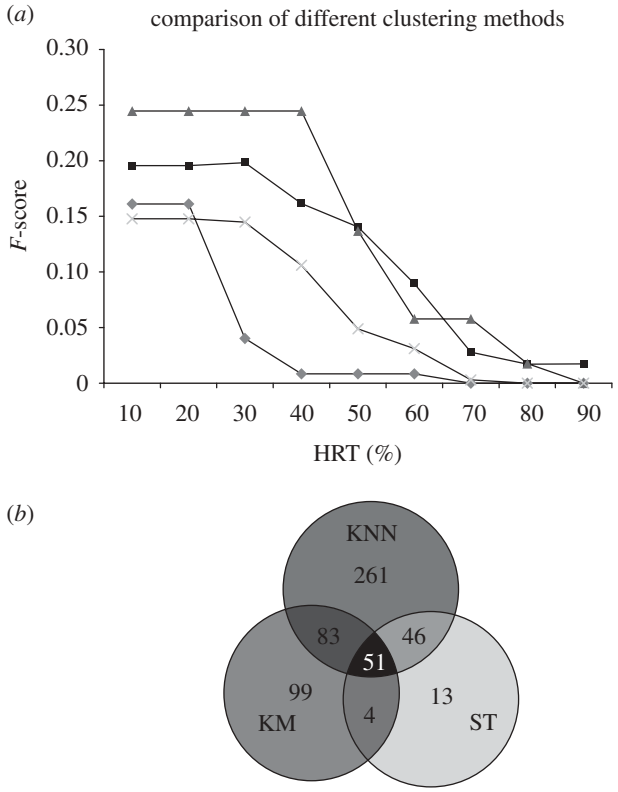

Figure 3. Comparison of different clustering methods. (*a*) *F*-scores for the four clustering methods (HIE (diamonds), KM (squares), KNN (triangles) and ST (crosses)) and their corresponding best-choice parameters ($K_{\mathrm{H}} = 1250$, $K_{\mathrm{KM}} = 1000$, $K_{\mathrm{KNN}} = 8$ and $\mathrm{SST} = 0.95$), over all HRTs in [10%, 90%]. (*b*) The number of predictions produced by the three clustering methods (KM, KNN and ST) for their corresponding best-parameter choices ($K_{\mathrm{KM}} = 1000$, $K_{\mathrm{KNN}} = 8$ and $\mathrm{SST} = 0.95$) at an HRT of 40 per cent.

a set of predictions for each of the four clustering methods. Since we predict a gene as cancer-related if a hit-rate (i.e. known cancer gene enrichment) of its cluster is above a given threshold, the critical factor is determining which HRT to use to make predictions. Currently, we do not know the entire set of cancer genes. Therefore, we expect that hit-rates of clusters will increase as new cancer genes are identified. Thus, we want to use a lower HRT to avoid too high stringency, but still large enough to make as many correct predictions as possible. Subsequently, we use HRT of 40 per cent, giving *F*-scores of 24.46 per cent, 16.16 per cent, 10.6 per cent and 0.8 per cent for KNN, KM, ST and HIE, respectively. These translate into precision of 31 per cent at recall of 20 per cent for KNN, precision of 31 per cent at recall of 11 per cent for KM, precision of 37 per cent at recall of 6 per cent for ST, and precision of 12.5 per cent at recall of 0.4 per cent for HIE (this is why we do not make predictions using HIE; also see below, and electronic supplementary material, figure S1). Note that relatively lower *F*-scores are not surprising for the following reasons. Since the number of known cancer genes will increase in the future, it is expected that precision, recall and *F*-scores will increase as well. Additionally, all of our predictions are already statistically significant; since we removed clusters for which cancer gene enrichment was likely to occur purely by chance, the number of possible predictions is automatically decreased, thus decreasing the *F*-scores. Finally, we are dealing with noisy and incomplete protein interaction datasets based on which we computed protein signatures.

Next, we assess the significance of the observed *F*-scores. We compare *F*-scores of real data with *F*-scores of randomized data (see §3), and we demonstrate that those for randomized data are much lower than those for real data. More specifically, real data *F*-scores of 24.46 per cent, 16.16 per cent, 10.6 per cent and 0.8 per cent for KNN, KM, ST and HIE, respectively, are higher than those of 8.39 per cent, 4.39 per cent, 5.71 per cent and 0.8 per cent for randomized data, respectively. Thus, we are confident in the accuracy of all clustering methods apart from HIE, for which *F*-scores for the real and randomized data are the same. Additionally, the performance of HIE drops dramatically at HRT of 30 per cent (figure 3*a*). For these reasons, we do not report our predictions for HIE. Instead, we report our predictions only for KNN, KM and ST.

With KNN, we produce 441 predictions, out of which 304 are 'new predictions' still unreported as known cancer genes in any of the databases and 137 are known cancer genes. With KM, we produce 237 predictions, out of which 163 are new predictions and 74 are known cancer genes. With ST, we produce 114 predictions, out of which 72 are new predictions and 42 are known cancer genes. The three clustering methods together make the total of 557 unique cancer gene predictions, out of which 399 are new predictions and 158 are known cancer genes. Thus, in total, we successfully recover 158 out of 679 known cancer genes; i.e. 23.3 per cent of them. Note that this relatively low recovery rate is expected and is the state of the art due to incompleteness and noise in the protein interaction network and known cancer gene dataset. Also note that our results outperform those reported by similar studies (Aragues *et al.* 2008), additionally verifying the superiority of our approach. In addition, we provide predictions confirmed by all three of these clustering methods, since these predictions could be considered of higher confidence. There are 51 such predictions, out of which 31 are new predictions and 20 are known cancer genes. The number of our cancer gene predictions for each of the three clustering methods, as well as their overlap, is presented in figure 3*b*. The existence of an overlap between different clustering methods demonstrates that it is difficult to choose a single 'best' clustering method. This further justifies our decision to report predictions produced by all three methods.

In table 1, we provide the list of 31 new predictions (as explained above) supported by all three clustering methods, together with *p*-values and hit-rates for their corresponding clusters. For a given prediction, the *p*-value (hit-rate) is the minimum *p*-value (maximum hit-rate) over all three clustering methods. The full list of predictions is available at http://www.ics.uci.edu/~bio-nets/predictions.xls.

Table 1. New cancer gene predictions supported by all three clustering methods and the *p*-values and hit-rates for their corresponding clusters. The predictions are sorted alphabetically with respect to gene names. If a gene is validated in the literature, the corresponding reference is shown ('PMID' denotes the PubMed ID).

gene	*p*-value	hit-rate (%)	reference (PMID)
ADRBK1	0.00051	83.33	18451066
ATF2	0.00001	57.14	18348191
CCL3L1	0.00248	62.50	15662971
CD247	0.00694	100.00	—
CD5	0.00014	83.33	19041428
CDC25B	0.03902	42.86	18635965; 17934831; 19028102
CRKL	0.00001	71.43	—
CSNK2A1	0.0003	71.43	15355908; 7513612
CSNK2B	0.01125	57.14	—
DAXX	0.00114	57.14	17952115; 17306074
HNF4A	0.001	71.43	18925631
KHDRBS1	0.00033	60.00	18394557; 17927519; 17621265
MAML3	0.00123	71.43	—
MAP3K14	0.01125	66.67	18434448
NCF1	0.00123	71.43	16753344
NR3C1	0.00001	57.14	18519667; 17167179; 14580768
PECAM1	0.00123	66.67	17875702; 17429142
PIK3R2	0.00022	71.43	15375573
PLCG1	0.00002	64.29	—
PLCG2	0.00183	57.14	—
POU2F1	0.00003	71.43	17273778; 15672409
PRKC1	0.01125	66.67	17990328; 17690741
PTPRC	0.00005	57.14	16818275; 16175399
PXN	0.00001	52.17	18990162; 18380937; 16040804
RAF1	0.00001	57.14	2018353; 18561318; 11389083
RAPIGA1	0.00248	62.50	—
RBL1	0.00114	44.44	17486638; 14666683
RUNX2	0.00292	43.75	18829534; 18755791
SHC1	0.00017	80.00	19055724; 18604176; 18273058
SQSTM1	0.01125	66.67	18931699; 17395976; 12700667
VDR	0.00051	83.33	19008093; 18849534; 18719092

4.2. Validation of our cancer gene predictions

4.2.1. Literature validation. To further demonstrate the correctness of our predictions, we perform literature validation. Owing to a large number of our new predictions, we perform *manual* literature validation only on 31 new predictions supported by all three clustering methods, successfully validating $24/31 = 77.4$ per cent of these predictions. Additionally, for *all* of our new predictions, we perform literature validation by automatic text mining using CiteXplore (Labarga *et al.* 2007). For $336/399 = 84.2$ per cent of them, this tool finds at least one article mentioning the protein of interest in the context of human cancer, thus further demonstrating the high prediction accuracy of our approach. Below, we provide our manually literature-validated predictions and the evidence of their involvement in cancer. The PubMed[6] IDs of references documenting each gene–cancer association are shown in table 1.

ADRBK1 (i.e. GRK2) plays a role in the growth of thyroid cancers reducing cell proliferation. ATF2 is involved in prostate cancer, breast cancer cell lines, hepatic and lung cancer, as well as melanoma and tumours of the nervous system. CCL3L1 is involved in brain tumorigenesis, and especially in the progression of glioblastoma. Expression of CD5 plays a role in the fate of tumour-specific T cells, making them unable to recognize and eliminate malignant cells. CDC25B overexpression has been observed in a significant number of human cancer cells, and in particular in early stages of development of intestinal and gastric cancers. CSNK2A1 plays a role in malignant progression and it encodes the catalytic subunit alpha of protein kinase CK2, where elevated CK2 activity is associated with malignant transformation of several tissue types, including lung. An inverse correlation is observed between proteins DAXX and c-Met in cancer cell lines and in metastatic breast cancer specimens; moreover, abnormal DAXX expression is observed in acute leukaemia.

Similarly, a role of HNF4A has been suggested in tumorigenesis and tumour development of hepatocellular carcinomas in mice. KHDRBS1 (i.e. P62 or SAM68) is overexpressed in human tumours and it is necessary for the survival of human lung adenocarcinoma cells; moreover, its pro-oncogenic role has been suggested and it has been identified as a modulator of tyrosine kinase activity and signalling requirement for mammary tumorigenesis and metastasis. An association has been identified between polymorphisms of MAP3K14 and rheumatoid arthritis, an inflammatory disease of the immune system associated with increased occurrence of cancers of the lymphatic system. Murine NCF1 gene has been linked to a pathway responsible for autoimmune disease and cancer. Expression of NR3C1 is linked to MT1-MMP in multiple tumour types, where invasion-promoting MT1-MMP is directly linked to tumorigenesis and metastasis; also, it has been identified as a gene with cancer-specific hypermethylation in colorectal tumours and as a cancer-associated gene with decreased expression. Bone marrow retention of acute myelogenous leukaemia cells depends on PECAM1 (CD31) coexpression level; also, its immunoreactivity closely parallels the different morphological steps of melanocytic tumour progression and the presence of histological parameters related to the aggressive behaviour.

Moreover, genetic alterations of ARHGAP family genes, including PIK3R2, lead to carcinogenesis through the dysregulation of Rho/Rac/Cdc42-like GTPases. POU2F subfamily members, including POU2F1, play a pivotal role for the FZD5 expression in undifferentiated human ES cells, foetal liver/spleen, adult colon, pancreatic islet and diffuse-type gastric cancer; also, a potential role of POU2AF1 in the deletion of 11q23 in chronic lymphocytic leukaemia

[6]http://www.pubmed.gov.

has been examined. PRKCI may serve as a molecular marker for metastasis and occult advanced tumour stages in oesophageal squamous cell carcinomas and it may represent novel tissue marker helpful in the differentiation of ductal and lobular breast cancers. PTPRC (CD45) antibodies may be candidates for immunotherapeutic approaches to the treatment of human NK and T cell lymphoma and its low expression and the poor response to CD3 have been recognized as markers that are able to identify rapid progress of gastric adenocarcinoma and the need for more aggressive therapeutic strategies. FYN and its related signalling partners, including PXN, are upregulated in prostate cancer; moreover, PXN plays an important role in controlling cell spread and migration and its overexpression correlates with the prognosis of some types of cancers, such as oesophageal cancer; furthermore, it functions as effector of GD3-mediated signalling, leading to malignant properties such as rapid cell growth and invasion.

Furthermore, RAF1 has been linked to human breast cancer as well as to urinary bladder cancer. The connection between the Rb family, including RBL1 (i.e. p107), and the TP53 family in skin carcinogenesis has been demonstrated; also, RBL1 plays a constitutive role in the progression of papillary carcinoma. RUNX2, in addition to being expressed in breast cancer cells, upregulates the cycle of metastatic bone disease, regulates genes related to progression of tumour metastasis and is involved in numerous disease processes, including postmenopausal osteoporosis and breast cancer. SHC1 has a role in endometriosis, a clinical condition that affects up to 10 per cent of women of reproductive age, characterized by the presence of endometrial tissues outside the uterine cavity and can lead to chronic pelvic pain, infertility and, in some cases, to ovarian cancer; also, it has been implicated in breast cancer. SQSTM1 (i.e. p62) is downregulated in hypoxia in carcinoma cells, and thus it might have a role in the regulation of hypoxic cancer cell survival responses; additionally, it is overexpressed in breast cancer. A role of polymorphisms in VDR has been demonstrated in cutaneous malignant melanoma and non-melanoma skin cancer risk, as well as in the development of sporadic prostate cancer and breast cancer.

4.2.2. Biological application and validation. Previous studies have hypothesized that cancer genes are negative regulators of melanogenesis in human cells (Halaban 2002). To examine the utility of our topological signatures for identifying cancer genes within biologically relevant datasets, we seek to identify cancer genes that are negative regulators of melanogenesis within our functional genomics dataset (Ganesan *et al.* 2008). We focus on our 695 most statistically significant genes that negatively regulate melanin production (the full list is available at http://www.ics.uci.edu/~bionets/mp_regulation.xls). To identify among negative regulators of melanin production those genes that are involved in cancer, we search within this dataset for our predictions produced by any of the three clustering algorithms. Four per cent of the negative regulators of melanogenesis, i.e. 27 genes, are identified as cancer gene candidates, out of which 14 are new predictions and 13 are known cancer genes. Of these 27 genes identified in this analysis, 85 per cent, i.e. 23 of them, are validated in the literature as cancer-associated genes (table 2).

Interestingly, 20 of these 27 genes are kinases, enzymes that are known to dynamically regulate the process of cellular transformation. Several of these kinases are known regulators of melanogenesis. BRAF, the top negative regulator of melanogenesis identified in our analysis, directly impacts melanin production by downregulating the transcriptional activity of MITF (Rotolo *et al.* 2005). Similarly, both MAPK1 (Yanase *et al.* 2001) and MAP3K1 (Hemesath *et al.* 1998) are known to directly impact MITF transcription. Also, among the cancer gene candidates is MERTK. Mutations in MERTK are responsible for retinitis pigmentosa, a hereditary cause of blindness characterized by aberrant melanin production in the retinal pigment epithelium (Gal *et al.* 2000). In addition to identifying known negative pigment regulators, our algorithm also identified several known uveal melanoma genes. Loss of heterozygosity in THRB is seen in hereditary uveal melanoma (Kos *et al.* 1999) while both MET and IGF1R play a role in uveal melanoma metastasis (Economou *et al.* 2008).

To further investigate the biological meaning of our results, we seek to determine whether our approach could identify sets of genes that are correlated with a specific cancer type. Note that complex wiring of cellular networks implies that the same gene can be involved in different biological processes and types of cancer (table 2). We find that 35 per cent of gene pairs from table 2 have signature similarities higher than the statistically significant threshold of 85 per cent (see §3). Next, we divide the set of genes from table 2 into subsets based on their involvement in a specific type of cancer. Interestingly, the percentage of statistically significantly signature-similar nodes is 40 per cent, 50 per cent and 50 per cent for breast, lung and digestive system cancer, respectively. This indicates that proteins involved in the same cancer type have more similar topologies in the PPI network than do the proteins involved in different cancer types. Finally, we examine the three uveal melanoma genes discussed above, THRB, MET and IGF1R. Interestingly, all three of these genes have very high signature similarities of above 92 per cent (figure 4), further indicating that genes involved in the same disease have very similar network neighbourhoods. The striking similarity of topological signatures of the three uveal melanoma genes is unlikely to occur purely by chance ($p \leq 0.018$; see §3). Since the three uveal melanoma proteins are at distances 2 and 3 in the PPI network from each other, no conclusion about their common biological characteristics could have been made from analysing less-constraining topological properties such as their direct neighbourhoods. Similarly, no such conclusion could have been made from comparing their sequences, since only MET and IGF1R have statistically significant BLAST sequence similarity, whereas protein pairs MET–THRB and IGF1R–THRB do not.

Together, these results demonstrate the utility of our topological similarity measure to discover new

Table 2. An siRNA-based screening approach identified 695 siRNAs that significantly stimulated pigment production. The corresponding genes were identified as negative regulators of melanin production (Ganesan *et al.* 2008). Twenty-seven out of these 695 genes, listed in the table, are identified as cancer gene candidates in this study. *Z*-scores quantify how significantly these genes stimulate pigment production. Hit-rates and *p*-values quantify cancer gene enrichments in the clusters formed for these genes and the statistical significance of observing given enrichments, respectively. Literature search and Gene Ontology database (The Gene Ontology Consortium 2000) were used to segregate these 27 genes into functional classes. PubMed search was used to identify known associations between these genes and human cancer. The specific cancer that each gene is associated with and references documenting each gene–cancer association are shown ('PMID' denotes the PubMed ID).

function	gene symbol	*Z*-score	hit-rate (%)	*p*-value	cancer association	reference (PMID)
kinases	*BRAF*	25.271	42.86	0.0675	multiple cancers	12068308
	MAP3K7	5.257	44.00	0.0001	breast cancer	18316610
	MAP2K1	5.243	42.86	0.0675	lung adenocarcinoma	18632602
	MST1R	4.193	57.14	0.0112	ovarian cancer	12915129
	HSPA1B	3.500	42.86	0.0675	hepatocellular carcinoma	18344806
	ABL1	3.479	75.00	0.0161	leukaemia	18704194
	ALK	3.300	42.86	0.0675	neuroblastoma, lung cancer, leukaemia	18923524
	FGFR1	3.279	42.86	0.0675	lung squamous cell carcinoma	18829480
	MAPK1	3.171	71.43	0.0012	breast cancer	18710790
	MAP3K1	2.800	66.67	0.0112	breast cancer	18437204
	KDR	2.743	42.11	0.00001	gastric adenocarcinoma	18609713
	ILK	2.386	57.14	0.0112	colon cancer	12771992
	FLT4	2.357	42.86	0.0675	osteosarcomas	18440723
	MERTK	2.279	42.86	0.0675	leukaemia	16675557
	MAP3K14	2.250	66.67	0.0112	prostate cancer	18752500
	IGF1R	2.207	71.43	0.0012	colon cancer	18636198
	MET	2.164	71.43	0.0012	multiple cancers	17992475
	JAK3	2.064	44.44	0.0.275	leukaemia	18559588
	MAP2K4	2.029	42.86	0.0675	pancreatic cancer	15623633
	MKNK2	7.657	40.00	0.0390	—	—
other	*CDH5*	2.100	45.46	0.0113	breast cancer	18316602
	USP2	3.221	66.67	0.0790	prostate cancer	15050917
	SMAD7	2.136	66.67	0.0062	leukaemia	18231913
	THRB	3.300	85.71	0.00002	pituitary carcinoma	18683837
	KPNB1	4.586	57.14	0.0112	—	—
	IVNSIABP	2.007	42.86	0.0675	—	—
	U2AF2	3.350	54.55	0.0018	—	—

biological knowledge without using any information external to PPI networks. Also, they demonstrate the utility of our approach to specifically identify cancer genes within phenotype specific, biologically relevant functional genomic datasets.

4.3. Comparison with other studies

Thus far, studies have been mainly focusing on examining global topological properties of cancer genes. In this light, Jonsson & Bates (2006) demonstrated greater connectivities and centralities of cancer genes compared with non-cancer genes, indicating an increased central role of cancer genes within the interactome. Unlike them, Goh *et al.* (2007) showed that the majority of disease genes were non-essential and that they did not show a tendency to code for hub proteins, thus indicating that the observed correlation between high degrees and disease genes was entirely due to the existence of essential genes within the disease gene class. We analysed several global network properties of both cancer genes and non-cancer genes in the context of our human PPI network. These network properties included the degree distribution, clustering spectrum and eccentricity. We observed no difference in the trends for cancer genes and in the trends for non-cancer genes with respect to these global network properties (data not shown). This clearly indicates the need to use more constraining local network properties such as node signature similarities.

To further demonstrate the strength of our method, we compare our results to those reported by Aragues *et al.* (2008) that also predict involvement of genes in cancer from protein interaction data. Unlike their approach that assumes that network neighbours of cancer genes are also involved in cancer (Aragues *et al.* 2008), we examine whether the genes that are involved in cancer have similar topological signatures without necessarily being adjacent in the network. We find that 96 per cent of known cancer gene pairs with signature similarities above the statistically significant threshold of 0.85 (see §3) are indeed not direct neighbours in the PPI network; more specifically, 3.88 per cent, 35.68 per cent, 48.53 per cent, 10.88 per cent, 1.02 per cent and 0.01 per cent of these pairs are at the shortest path distance of 1, 2, 3, 4, 5 and 6, respectively. Note that in addition to the interaction data alone, Aragues *et al.* (2008) also use differential expression data, as well as structural and functional properties of cancer genes. We demonstrate that our method, based solely on network topology, outperforms the method of Aragues *et al.* (2008) even when they use

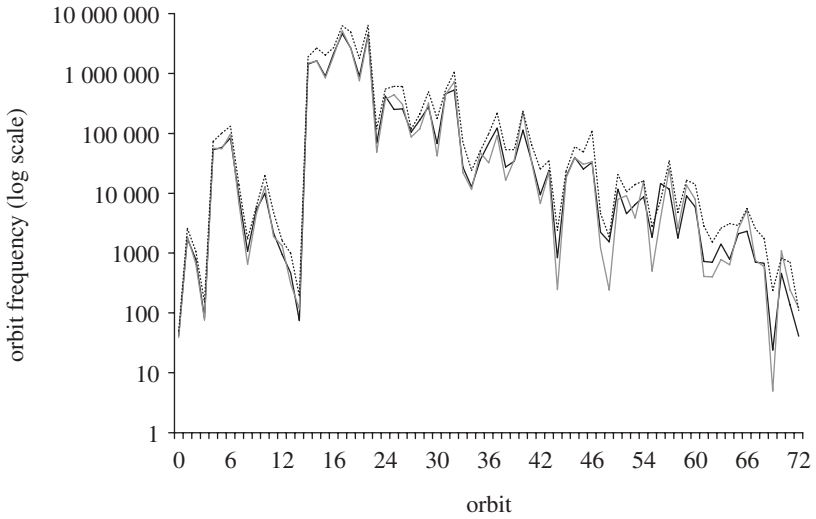

Figure 4. Signature vectors of three uveal melanoma genes (THRB (black), MET (grey) and IGF1R (dotted)). Their signature similarities are very high, above 92 per cent.

additional biological information. They report that using solely interaction data results in precision of 54 per cent at recall of 1 per cent, which corresponds to an F-score of 16.87 per cent. The precision of their integrative approach when multiple data types are used ranges from 23 per cent at recall of 15 per cent to 73 per cent at recall of 1 per cent, corresponding to the range of F-scores from 18.15 per cent down to 1.97 per cent. Note that the quality of our predictions compares favourably even with their integrative approach, since it ranges from precision of 31 per cent at recall of 20 per cent to precision of 86 per cent at recall of 1 per cent (electronic supplementary material, figure S1), corresponding to the range of F-scores from 24.46 per cent down to 1.98 per cent.

Thus, using PPI network topology only, we clearly outperform the best results that they obtained by integration of PPI network and other data types. A reason for this could be that Aragues *et al.* (2008) relied only on the direct neighbours of genes when implying their involvement in cancer while the signatures are a more precise measure of network structure capturing up to 4-deep neighbourhoods of proteins in PPI networks. Aragues *et al.* (2008) also performed a literature search validating 60 per cent of their predictions. We obtain a higher literature validation hit-rate of 77.4 per cent. Furthermore, unlike Aragues *et al.* (2008), we provide biological application and validation of our approach to melanogenesis-related functional genomics data.

There are several potential limitations of our approach. First, the protein interaction network data from which protein signatures are computed are noisy and incomplete, thus affecting the resulting signature similarities, clusters and predictions. However, we previously tested our method both on high- and low-confidence PPI networks and showed its robustness to noise in PPI network data (Milenković & Pržulj 2008). Second, graphlet signature similarity is currently the most constraining measure of topological similarity of nodes in a network. Thus, it might be too stringent to detect weak PPI network similarities between proteins that share biological function, despite their network topological differences. However, we still outperform studies that use less-constraining global network properties such as direct neighbours of nodes (Aragues *et al.* 2008). Third, the set of known cancer genes is still incomplete, so it might be difficult to identify all new cancer gene candidates. Note, however, that this is not a limitation of our method, but of the incompleteness of data sets. Also note that we still outperform other related approaches (Aragues *et al.* 2008). Finally, we do not provide explicit experimental validation of our cancer gene predictions. However, we identify our cancer gene predictions in the melanoma-related functional genomics data set predicted to contain them. Additionally, we validate about 80 per cent of our predictions in the literature. Thus, despite potential limitations of our method and the fact that we might miss some cancer gene candidates, we are still confident in our predictions.

5. CONCLUSION

We address the important challenge of identifying the relationship between PPI network topology and disease. Based solely on topological signatures of proteins in PPI networks, we are able to identify statistically significant and biologically relevant cancer gene candidates. We determine that this approach can specifically identify cancer genes within systems-level functional genomics datasets predicted to contain them. Thus, our method can be used to probe biologically relevant datasets and uncover novel relationships between cancer genes and specific cellular phenotypes. Interpretation and analysis of the results of systems-level functional genomics analysis is limited by the high false-negative rate of this analysis. Using protein interaction network topology to analyse functional genomics datasets can potentially uncover specific

molecular pathways that regulate a given biological phenotype.

It would be interesting to examine whether the topological signatures could be used to further distinguish between different types of cancer or even between different disease classes. A potential improvement of our approach would be to integrate with proteins' PPI network topological signatures other types of biological data such as gene expression, protein structure and functional information. By validating our predictions, we provide clear evidence that PPI network structure around cancer genes is different from the structure around non-cancer genes. Understanding the fundamental laws underlying this observation is an open research problem that represents a promising avenue towards understanding and eventually treating complex diseases.

We thank Oleksii Kuchaiev for useful discussions and suggestions. We also thank Michael A. White for his guidance during the completion of the genome-wide siRNA screen. This project was supported by the NSF CAREER IIS-0644424 grant, a UCI CCBS 2008 Opportunity Award and a grant from Outrun the Sun, Inc.

REFERENCES

Aragues, R., Sander, C. & Oliva, B. 2008 Predicting cancer involvement of genes from heterogeneous data. *BMC Bioinformatics* **9**, 172. (doi:10.1186/1471-2105-9-172)

Barabási, A. & Oltvai, Z. N. 2004 Network biology: understanding the cell's functional organization. *Nat. Rev.* **5**, 101–113. (doi:10.1038/nrg1272)

Brun, C., Herrmann, C. & Guénoche, A. 2004 Clustering proteins from interaction networks for the prediction of cellular functions. *BMC Bioinformatics* **5**, 95. (doi:10.1186/1471-2105-5-95).

Economou, M. A., All-Ericsson, C., Bykov, V., Girnita, L., Bartolazzi, A., Larsson, O. & Seregard, S. 2008 Receptors for the liver synthesized growth factors IGF-1 and HGF/SF in uveal melanoma: intercorrelation and prognostic implications. *Acta Ophthalmol.* **4**, 20–5. (doi:10.1111/j.1755-3768.2008.01182.x)

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. & Stratton, M. R. 2004 A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183. (doi:10.1038/nrc1299)

Gal, A., Li, Y., Thompson, D. A., Weir, J., Orth, U., Jacobson, S., Apfelstedt-Sylla, E. & Vollrath, D. 2000 Mutations in MERTK, the human orthologue of the rcs rat retinal dystrophy gene, cause retinitis pigmentosa. *Nat. Genet.* **26**, 270–271. (doi:10.1038/81555)

Ganesan, A. *et al.* 2008 Genome-wide siRNA-based functional genomics of pigmentation identifies novel genes and pathways that impact melanogenesis in human cells. *PLoS Genet.* **4**, e1000298. (doi:10.1371/journal.pgen.1000298)

Goh, K., Cusick, M. E., Valle, D., Childs, B., Vidal, M. & Barabási, A.-L. 2007 The human disease network. *Proc. Natl Acad. Sci. USA* **104**, 8685–8690. (doi:10.1073/pnas.0701361104)

Guerrero, C., Milenković, T., Pržulj, N., Jones, J. J., Kaiser, P. & Huang, L. 2008 Characterization of the yeast proteasome interaction network by QTAX-based tagteam mass spectrometry and protein interaction network analysis. *Proc. Natl Acad. Sci. USA* **105**, 13 333–13 338. (doi:10.1073/pnas.0801870105)

Halaban, R. 2002 Pigmentation in melanomas: changes manifesting underlying oncogenic and metabolic activities. *Oncol. Res.* **13**, 3–8.

Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D. & McKusick, V. A. 2002 Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **30**, 52–55. (doi:10.1093/nar/30.1.52)

He, X. & Zhang, J. 2006 Why do hubs tend to be essential in protein networks? *PLoS Genet.* **2**, e88. (doi:10.1371/journal.pgen.0020088)

Hemesath, T. J., Price, E. R., Takemoto, C., Badalian, T. & Fisher, D. E. 1998 MAP kinase links the transcription factor Microphthalmia to c-Kit signalling in melanocytes. *Nature* **391**, 298–301. (doi:10.1038/34681)

Ideker, T. & Sharan, R. 2008 Protein networks in disease. *Genome Res.* **18**, 644–652. (doi:10.1101/gr.071852.107)

Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. 2001 Lethality and centrality in protein networks. *Nature* **411**, 41–42. (doi:10.1038/35075138)

Jonsson, P. & Bates, P. 2006 Global topological features of cancer proteins in the human interactome. *Bioinformatics* **22**, 2291–2297. (doi:10.1093/bioinformatics/btl390)

Kanehisa, M. & Goto, S. 2000 KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30. (doi:10.1093/nar/28.1.27)

King, A. D., Pržulj, N. & Jurisica, I. 2004 Protein complex prediction via costbased clustering. *Bioinformatics* **20**, 3013–3020. (doi:10.1093/bioinformatics/bth351)

Kos, L., Aronzon, A., Takayama, H., Maina, F., Ponzetto, C., Merlino, G. & Pavan, W. 1999 Hepatocyte growth factor/scatter factor-MET signaling in neural crest-derived melanocyte development. *Pigment Cell Res.* **12**, 13–21. (doi:10.1111/j.1600-0749.1999.tb00503.x)

Krishnan, M. N. *et al.* 2008 RNA interference screen for human genes associated with West Nile virus infection. *Nature* **455**, 242–245. (doi:10.1038/nature07207)

Kuchaiev, O., Milenković, T., Memišević, V., Hayes, W. & Pržulj, N. 2009 Topological network alignment uncovers biological function and phylogeny. (http://arXiv.org/abs/0810.3280v2)

Labarga, A., Valentin, F., Andersson, M. & Lopez, R. 2007 Web Services at the European Bioinformatics Institute. *Nucleic Acids Res.* **35**, W6–W11. (doi:10.1093/nar/gkm291)

Milenković, T. & Pržulj, N. 2008 Uncovering biological network function via graphlet degree signatures. *Cancer Informatics* **4**, 257–273.

Peri, S. *et al.* 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13**, 2363–2371. (doi:10.1101/gr.1680803)

Pržulj, N. 2007 Biological network comparison using graphlet degree distribution. *Bioinformatics* **23**, e177–e183. (doi:10.1093/bioinformatics/btl301)

Pržulj, N. & Milenković, T. In press. Computational methods for analyzing and modeling biological networks. In *Biological data mining* (eds J. Chen & S. Lonardi). CRC Press.

Pržulj, N., Corneil, D. G. & Jurisica, I. 2004*a* Modeling interactome: scale-free or geometric? *Bioinformatics* **20**, 3508–3515. (doi:10.1093/bioinformatics/bth436)

Pržulj, N., Wigle, D. & Jurisica, I. 2004*b* Functional topology in a network of protein interactions. *Bioinformatics* **20**, 340–348. (doi:10.1093/bioinformatics/btg415)

Radivojac, P., Peng, K., Clark, W. T., Peters, B. J., Mohan, A., Boyle, S. M. & Mooney, D. S. 2008 An integrated approach to inferring gene–disease associations in humans. *Proteins* **72**, 1030–1037. (doi:10.1002/prot.21989)

Rotolo, S., Diotti, R., Gordon, R. E., Qiao, R. F., Yao, Z., Phelps, R. G. & Dong, J. 2005 Effects on proliferation and melanogenesis by inhibition of mutant braf and expression of wild-type ink4a in melanoma cells. *Int. J. Cancer* **115**, 164–169. (doi:10.1002/ijc.20865)

Safran, M. *et al.* 2002 Genecards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics* **18**, 1542–1543. (doi:10.1093/bioinformatics/18.11.1542)

Schwikowski, B. & Fields, S. 2000 A network of protein–protein interactions in yeast. *Nat. Biotechnol.* **18**, 1257–1261. (doi:10.1038/82360)

Sharan, R. & Ideker, T. 2006 Modeling cellular machinery through biological network comparison. *Nat. Biotechnol.* **24**, 427–433. (doi:10.1038/nbt1196)

Sharan, R., Ulitsky, I. & Shamir, R. 2007 Network-based prediction of protein function. *Mol. Syst. Biol.* **3**, 88. (doi:10.1038/msb4100129).

Silva, J. M., Marran, K., Parker, J., Silva, J., Golding, M., Schlabach, M. R., Elledge, S. J., Hannon, G. J. & Chang, K. 2008 Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science* **319**, 617–620. (doi:10.1126/science.1149185)

Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A. & Tyers, M. 2006 BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535–D539. (doi:10.1093/nar/gkj109)

The Gene Ontology Consortium. 2000 Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29. (doi:10.1038/75556)

Whitehurst, A.W. *et al.* 2007 Synthetic lethal screen identification of chemosensitizer loci in cancer cells. *Nature* **446**, 815–819. (doi:10.1038/nature05697)

Xu, R. & Wunsch, D. 2005 Survey of clustering algorithms. *IEEE Trans. Neural Netw.* **16**, 645–678. (doi:10.1109/TNN.2005.845141)

Yanase, H., Ando, H., Horikawa, M., Watanabe, M., Mori, T. & Matsuda, N. 2001 Possible involvement of ERK1/2 in UVA-induced melanogenesis in cultured normal human epidermal melanocytes. *Pigment Cell Res.* **14**, 103–109. (doi:10.1034/j.1600-0749.2001.140205.x)

Yidirim, M. A., Goh, K.-I., Cusick, M. E., Barabási, A.-L. & Vidal, M. 2007 Drug–target network. *Nat. Biotechnol.* **25**, 1119–1126. (doi:10.1038/nbt1338).

Zotenko, E., Mestre, J., O'Leary, D. P. & Przytycka, T. M. 2008 Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comp. Biol.* **4**, e1000140. (doi:10.1371/journal.pcbi.1000140)