Network propagation: a universal amplifier of genetic associations

Lenore Cowen[1], Trey Ideker[2], Benjamin J. Raphael[3] and Roded Sharan[4]

Abstract | Biological networks are powerful resources for the discovery of genes and genetic modules that drive disease. Fundamental to network analysis is the concept that genes underlying the same phenotype tend to interact; this principle can be used to combine and to amplify signals from individual genes. Recently, numerous bioinformatic techniques have been proposed for genetic analysis using networks, based on random walks, information diffusion and electrical resistance. These approaches have been applied successfully to identify disease genes, genetic modules and drug targets. In fact, all these approaches are variations of a unifying mathematical machinery — network propagation — suggesting that it is a powerful data transformation method of broad utility in genetic research.

Nodes
The objects modelled by a network. In biological networks, nodes can represent proteins, genes, metabolites, RNA molecules, or even diseases and phenotypes.

Edges
Relationships between pairs of nodes in a network, for example, molecular interactions between the genes or proteins that correspond to these nodes. Two nodes sharing an edge are said to be adjacent, neighbours, or directly connected by it.

[1]*Department of Computer Science, Tufts University, Medford, Massachusetts 02155, USA.*
[2]*University of California San Diego, La Jolla 92093, California, USA.*
[3]*Department of Computer Science, Princeton University, Princeton, New Jersey 08540, USA.*
[4]*Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel.*

Correspondence to R.S.
roded@post.tau.ac.il.

doi:10.1038/nrg.2017.38
Published online 12 Jun 2017

Genomic technologies have spawned numerous research efforts to decipher the molecular basis of disease, producing ever-increasing amounts of 'omics' data. Identifying significant patterns in these data has become a central challenge in genetics and medicine and involves dealing with noisy and incomplete observations, which requires the integration of multiple data types within a single analysis framework. A promising approach to overcome these hurdles and to boost the signal-to-noise ratio is to analyse the data in the context of molecular networks, be they physical, genetic, co-expression or other networks[1].

A molecular network model consists of nodes, representing molecules such as proteins, and edges that connect the nodes, representing pairwise relationships between the corresponding molecules, such as protein–protein interactions (PPIs). It serves as a convenient computational model for molecular data owing to its generality, representation simplicity and power to detect complex patterns, such as clusters, which cannot be readily gleaned from the pairwise data[2].

In the context of genetic association, early network analysis methods relied on the principle of 'guilt by association', based on observations that a gene or protein shares many molecular and phenotypic characteristics with its direct interactors[3]. Generalizing beyond direct connections to the concept of a local network neighbourhood has led to a plethora of methods for finding clusters or modules in a network[4]; however, module-based approaches were found to be less effective than 'direct' methods at associating proteins with their functional roles[5,6].

Recently, a new group of methods accounting for the global structure of the network has emerged as the state-of-the-art in genetic association[7,8]. At the heart of these methods lies the common paradigm of network propagation, which amplifies a biological signal based on the assumption that genes underlying similar phenotypes tend to interact with one another[9]. To this end, prior information associating genes with a phenotype of interest (for example, membership in a biological process or the presence of polymorphisms linked to a disease) is superimposed on the nodes of the network. The information is then propagated through the edges to nearby nodes in an iterative manner for a fixed number of steps or until convergence. The final value of a node is influenced by the values of its direct network neighbours, which in turn are affected by their neighbours, and so on. New nodes that were not included in the prior information can nonetheless be associated with the phenotype, with their propagation values reflecting proximity to the prior nodes (FIG. 1).

Because this propagation paradigm is very powerful, it has been discovered and re-discovered in numerous fields under different guises[10–14]. For example, graph theoreticians investigate random walks on graphs[11,15]; the data science community applies variants of the Google PageRank search algorithm[12,16]; statistical physicists study heat diffusion processes[17,18]; electrical engineers compute minimum energy states within an electrical circuit[14,19]; and the machine-learning community considers different forms of graph kernels[20]. In biology, these different formulations of network propagation have been used for various

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

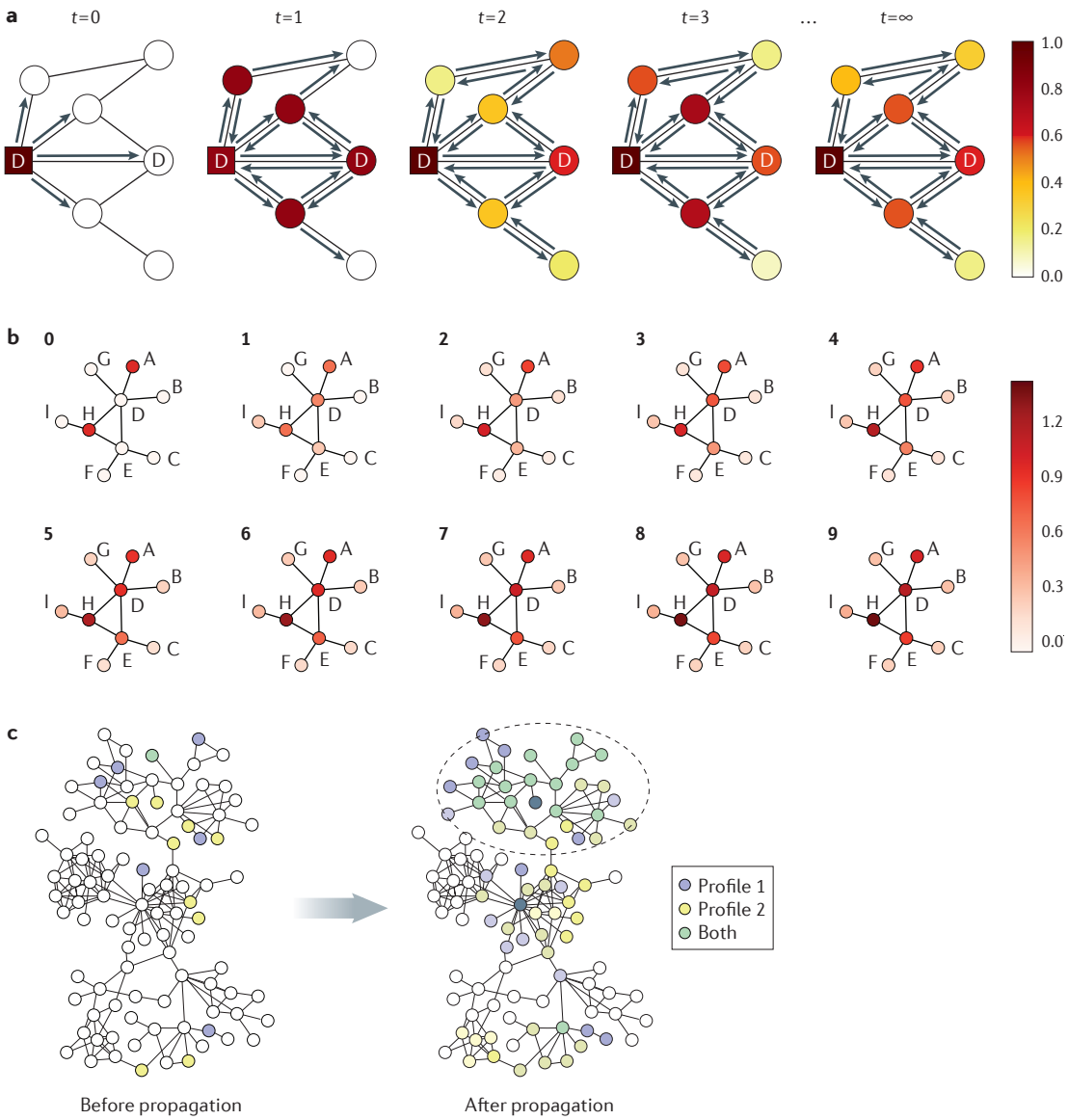

Figure 1 | **Schematic illustration of network propagation. a** | A step-by-step demonstration of network propagation. The propagation process is depicted at different time points until convergence (steady-state ($t = \infty$)). Arrows depict the direction of the flow or walk. Nodes are colour-coded according to the amount of flow that they receive. D indicates nodes that are known (square node) or that are predicted (circular node) to be associated with a disease phenotype. **b** | Example network with initial high scores for two of nine nodes (step 0, nodes A and H; score shown by colour bar). These scores are allowed to propagate over stepwise iterations 0–9; note that convergence is reached by approximately step 5 and thus the colours do not change markedly in subsequent steps. **c** | Illustration of a biological network with gene scores before and after propagation, performed independently for two data sets (profile 1 and profile 2). Propagation results in greater concordance between the data sets, as is evident from the greater number of green nodes (dashed oval). Part **c** is adapted with permission from REF. 89, Macmillan Publishers Limited.

Network propagation
A family of stochastic processes that trace the flow of information through a network over time.

Random walks
Mathematical formalization of the paths resulting from taking successive random steps. Classical examples of random walks are Brownian motion, the fortune of a gambler flipping a coin or fluctuations of the stock market. In the context of networks, a random walk typically describes a process in which a 'walker' moves from one node to another with a probability that is proportional to the weight of the edge connecting the nodes.

Kernels
Symmetric similarity functions with the property that one can assign vectors (in some abstract space) to its arguments such that the similarity of two elements is the dot-product between their corresponding vectors.

Disease module
A network module, the member genes of which are associated with a particular disease.

applications, including gene function prediction[6,21], module discovery[22], disease characterization[23,24] and drug target prediction[25].

In this Review, we discuss these different methods and their applications, focusing on network propagation as a unifying paradigm. We start by describing network propagation and outlining its suitability for network analysis. We then review early applications of this approach to biological research, primarily to the problem of protein function prediction. The focal part of this Review describes the applications of network propagation to analyse human diseases, including gene prioritization, disease module discovery, disease subtyping and drug target prediction.

Why network propagation?

Often in disease genetics, we are given a list of genes that previous studies have shown are associated with a disease (for example, by genome-wide association studies), and we wish to prioritize other genes that

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

may be associated with that disease. Given a network of interactions among these genes (such as a previously measured PPI network), we invoke the principle that disease-related genes are more likely to have biological interactions with each other than with randomly chosen genes.

A straightforward analysis approach might be to predict that all the direct neighbours of disease genes in the network are also disease genes[26] (FIG. 2a; left panel). However, such a naive approach would potentially introduce false predictions (false positives) that are connected to disease genes by irrelevant edges; it would also miss genes (false negatives) that do not directly interact with known disease genes, even if such genes are well connected to the known genes through multiple longer paths (FIG. 2a,b). To address this issue, one could examine longer paths in the network and could define the distance between pairs of genes by the length of the shortest path between them. One could then prioritize new genes on the basis of their distance to the prior list[27] (FIG. 2a; middle panel). However, the difficulty with this approach is that many genes will be near disease genes owing to the 'small world' property of most biological networks[28]; that is, the property that most nodes can be reached from every other node in a small number of steps. Thus, such an approach might return many false-positive genes that connect to disease genes through paths that contain irrelevant or erroneous interactions (FIG. 2a,b).

Network propagation offers a more refined approach by simultaneously considering all possible paths between genes (FIG. 2a; right panel). The application of network propagation to gene ranking can thus overcome some of the difficulties associated with shortest path-based approaches (FIG. 2a,b). Specifically, potentially spurious predictions (false positives) that are supported by a single (shortest) path are down-weighted, and true causal genes that are potentially missed, even though they are well connected to the prior list (false negatives), are promoted.

To illustrate the power of network propagation in a different application, we consider the problem of predicting genes that when somatically mutated contribute to the growth and development of cancer. Given the frequency of somatic mutations in genes across a cohort of patients with cancer, one approach to distinguish cancer-causing 'driver' genes from randomly mutated 'passenger' genes is to rank the genes according to their mutation frequency across patients[29]. However, ranking genes by mutation frequency alone performs poorly because driver and passenger genes can be mutated at similar frequencies, particularly in cohorts of a modest size[30]. However, by 'smoothing' the mutation frequencies across the network using a propagation process, the resulting predictions become highly significant because cancer driver genes tend to cluster in the network[31] (FIG. 2c).

A unifying formulation

Network propagation describes multiple techniques that follow the same underlying strategy (BOX 1). Suppose we have a partially labelled network, in which the labels on the nodes correspond to genes that are known to have certain molecular or phenotypic properties that are likely to be shared with genes in their local network neighbourhood. Consider the following propagation or diffusion process (FIG. 1a): a label is replaced with a certain amount of fluid, and at each time step the fluid flows to the neighbours of the corresponding node, either in equal proportions (in an unweighted network) or proportional to its edge weight. Then, halt this process after a small number of steps, when most of the fluid is still close to the original (labelled) nodes. We estimate that an unlabelled node has the property in question with a probability proportional to the amount of fluid that reached it from the labelled nodes. Despite its conceptual simplicity, the propagation process has several desired properties that account for the global structure of the network: first, it can score distant nodes that are not direct neighbours of the labelled nodes, as fluid will also reach those nodes; second, the process automatically adjusts for local connectivity, in the sense that nodes that are well connected through many short paths to the labelled node will receive more fluid; and third, the process automatically deprecates paths that go through highly connected or hub nodes, because in such paths the fluid will diffuse through to the many neighbours of the hub and so each neighbour will only receive a small share.

As formulated above, the propagation process is not informative if it is run for too long, because the fluid will eventually spread out over the whole network and will no longer capture the local neighbourhood of the labelled nodes. As an alternative to halting the process after a small number of steps, the random walk with restart (RWR) formulation adds a reset parameter: at each step of the propagation process, with some fixed probability (the restart value), rather than the fluid at a node continuing to propagate, it is returned to the original source node. This restart value serves as a damping factor on long walks, so that the fluid that reaches a node exponentially decreases on the basis of its distance from the source. This version has the advantage that the propagation process can be run to a steady state, and that the diffusion is confined to the local neighbourhood even at steady state (FIG. 1b).

The description thus far assumes that propagation to neighbours occurs at discrete time steps. However, some formulations of network propagation, particularly in physics, model a continuous fluid flow over time. In a formulation of network propagation in which the process is halted after a fixed number of discrete time steps, the amount of fluid that ends up at all network nodes can be computed by direct simulation. In formulations in which the underlying propagation process is continuous and/or in which one seeks to measure the amount of fluid at steady state, the result can be computed using classical matrix algebra (BOX 1).

We note that network propagation is inherently directed, as the fluid flows away from the source nodes; thus, the propagation result at node i when diffusing from node j may differ from the propagation result at node j when diffusing from node i. However, the underlying network over which the propagation process is run

False positives
Error in prediction whereby negative examples are predicted to be positive. For example, when predicting disease genes, a false positive would correspond to a non-disease gene that is wrongly predicted to be disease-related.

False negatives
Error in prediction whereby positive examples are predicted to be negative. For example, when predicting disease genes, a false negative would correspond to a disease gene that is missed and predicted to be unrelated.

Edge weight
An abstract measure of the 'strength' of the connection between a pair of nodes in a network, typically represented as a real number between 0 and 1.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

is typically assumed to be undirected[32], either because it models undirected relationships (such as co-complex membership) or because of missing information about edge directions (which is the case for most PPIs[33]). The asymmetry in the propagation process has motivated the development of several variants of network propagation to symmetrize this process, yielding kernels that can be easily incorporated into standard machine-learning pipelines. Nevertheless, some networks are naturally directed, such as signalling and regulatory networks. Although the basic network propagation paradigm can be adapted to the directed case

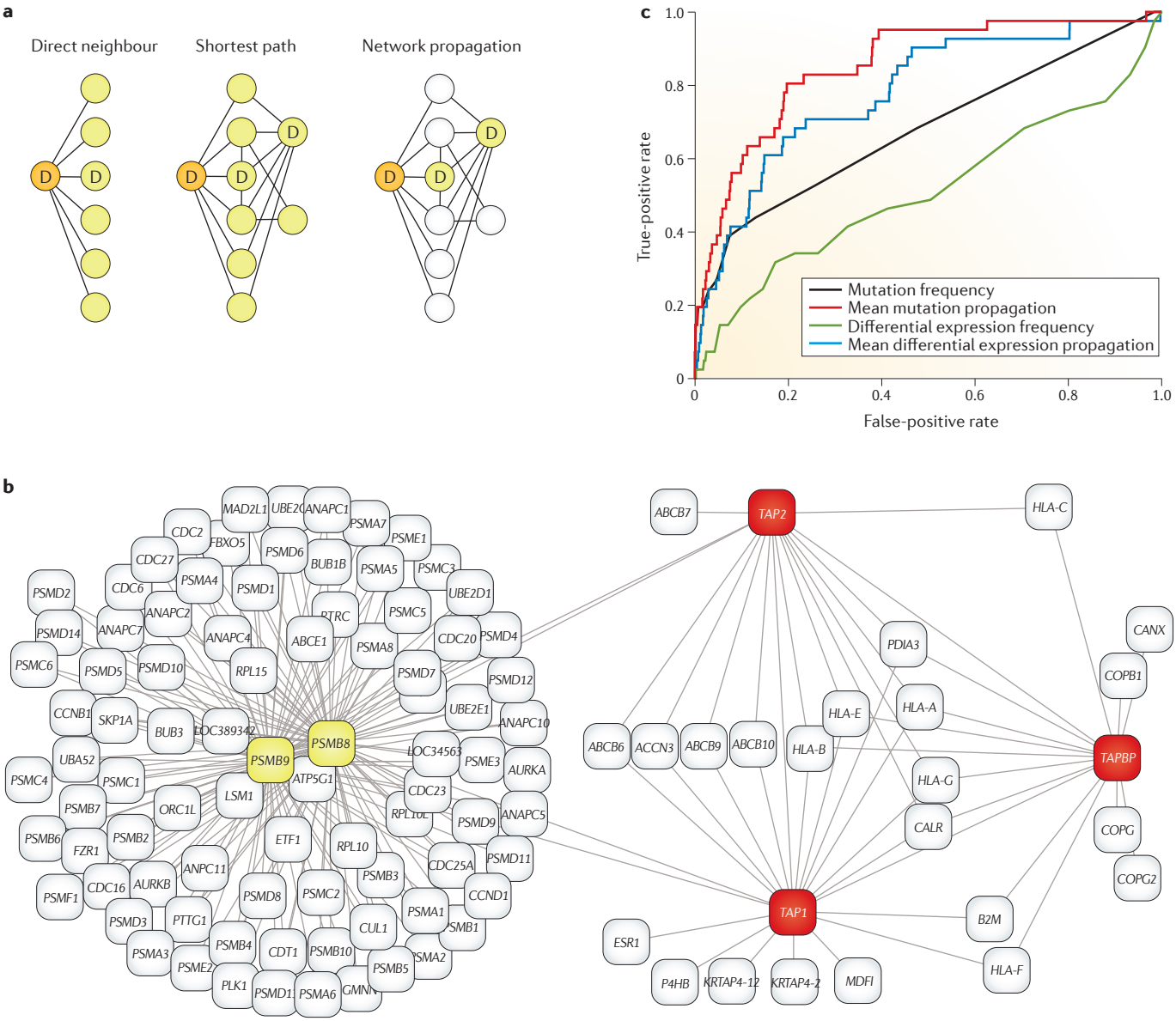

Figure 2 | **Network propagation for discovery and prioritization of disease genes. a** | A schematic example in which a single disease gene (orange) is used to identify additional disease-related genes; known disease genes are denoted by D. Predicting the involvement of the direct interactors of this gene (yellow; left panel) leads to many false positives, as well as to a false negative (shown in the two other panels). Looking at more distant neighbours that are up to two steps away (yellow; middle panel) again introduces many false positives. Network propagation overcomes these problems by simultaneously considering all paths between genes (yellow; right panel). **b** | A real example of a protein interaction network that is associated with bare lymphocyte syndrome type 1. Propagation of the signal from any of the three known disease genes (red) ranks the other known disease genes very highly, owing to the many paths between them. Genes in yellow are ranked highly by alternative network analysis methods (which consider direct neighbours or shortest paths); however, these are false positives. **c** | Receiver operating characteristic (ROC) curves for recovering known cancer genes defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG) glioblastoma pathway[101]. Performance over a set of 591 glioblastoma samples is shown for four different gene rankings according to differential mRNA expression between the tumour and normal samples (green)[102], somatic mutation frequencies in tumours (black)[102], network-propagated differential mRNA expression (mean across samples; blue) and network-propagated somatic mutations (mean across samples; red). Both network propagation variants considerably outperform their frequency-based counterparts (compare the blue curve to the green curve, and the red curve to the black curve). Part **b** is reproduced with permission from REF. 73, Elsevier. Part **c** is reproduced from REF. 31, Elsevier.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Box 1 | The mathematics of network propagation

Network propagation encompasses related and, in certain cases, mathematically equivalent techniques, including random walks on a graph, diffusion processes on a graph and current computations in electric networks (see the table). The starting point is a vector $p_0(v)$ of scores on genes representing our prior knowledge or experimental measurements. For example, we could set $p_0(v) = 1$ for known disease genes and $p_0(v) = 0$ for all other genes. Alternatively, we could set $p_0(v)$ to represent some measure of confidence in the role of v in a disease, for example, its frequency of somatic mutations when studying cancer cohorts. Conceptually, one can think of $p_0(v)$ as an amount of heat, fluid or information that diffuses (or flows) over the edges of the network. At each time point k, the amount of information at each node v depends on the sum of the information at the neighbouring (adjacent) nodes $N(v)$ at time $k-1$, in proportion to the weights on the corresponding edges, according to the following equation:

$$p_k(v) = \sum_{u \in N(v)} p_{k-1}(u) w(u, v) \qquad (1)$$

where $w(u,v)$ is the (normalized) weight or the confidence of the interaction between u and v. If we run this process for k steps, then the values in the resulting vector $p_k(v)$ give us a ranking of each node. When k is small, the ranking is close to the initial distribution $p_0(v)$, but when k is large, the information diffuses away from the initial distribution and reflects the network topology.

The propagation process described in Equation 1 can be written in matrix notation as follows:

$$p_k = W p_{k-1} \qquad (2)$$

where W is a normalized version of the adjacency matrix of the network of interest. Repeated iteration of this equation yields $p_k = W^k p_0$, where p_0 represents our initial, or prior, information on genes. If W is a stochastic matrix, that is, its columns sum to 1, this process is equivalent to a random walk on the network, where a walker traverses the nodes, each time moving to a random neighbour of the present position with a probability given by (the transpose of) W. Alternatively, we can view the edges as representing conductance in an electric network with some designated source and target. If one unit of current flows through the source, then the amount of current flowing through any edge is the frequency with which a random walker traverses that edge on the way from the source to the target.

Another version of the propagation process is the random walk with restart (RWR; also known as insulated diffusion and personalized PageRank):

$$p_k = \alpha p_0 + (1-\alpha) W p_{k-1} \qquad (3)$$

where the parameter α describes the trade-off between prior information and network smoothing. When the network is connected and the eigenvalues of W are at most 1 in absolute value, then this process can be shown to converge to a steady-state distribution:

$$p = \alpha (I - (1-\alpha) W)^{-1} p_0 \qquad (4)$$

Different variants may use different ways of defining W based on the adjacency matrix A of the network (which could be weighted or unweighted) and the diagonal degree matrix D, the diagonal entries of which hold the node degrees and all other entries are 0. The random walk above uses $W = AD^{-1}$. Other approaches set W to $D^{-1/2}AD^{-1/2}$, which also satisfies the convergence conditions.

In both cases, the final ranking can be obtained from the initial ranking by matrix multiplication: if we denote by p either the steady-state distribution or the diffusion at some time point k, then $p = Sp_0$ for some appropriately defined matrix S. This matrix can be interpreted as a (potentially asymmetric) similarity matrix, in which each entry S_{ij} gives the amount of information propagated to node i, given that the initial ranking p_0 is an elementary vector with 1 at entry j and 0 elsewhere.

Furthermore, if S is symmetric and positive semi-definite, then S defines a kernel. For example, the diffusion kernel is the continuous-time analogue of RWR, where $S = e^{-\alpha W}$ and $W = D - A$ is called the network's Laplacian matrix. The kernel framework for interpreting S is a useful one, because kernels can be easily plugged into general machine-learning algorithms for classification and regression.

The propagation variants described above can be readily generalized to weighted networks, but the situation with regard to directed networks is more complex. Although PageRank and personalized PageRank were developed and studied in a directed setting, the closed forms for the RWR steady-state and kernel formulations given above only hold for the undirected case. Chung[99] defined a natural analogue of the Laplacian matrix for directed graphs and used it to study the rate of convergence of random walks in the directed case. Other strategies for dealing explicitly with directed edges are surveyed in REF. 100.

Name	Similarity matrix	Weight normalization	Equivalent methods
Random walk	W^k	$W = AD^{-1}$	Electric network
Random walk with restart	$\alpha(I - (1-\alpha)W)^{-1}$	$W = AD^{-1}$; $W = D^{-1/2}AD^{-1/2}$	Insulated diffusion; personalized PageRank
Diffusion kernel	$e^{-\alpha W}$	$W = D - A$	Heat kernel

k denotes the number of time steps; A denotes the adjacency matrix, which could be weighted or unweighted; D denotes the diagonal degree matrix; α is the smoothing parameter

Adjacency matrix
A matrix representation of a network such that the *(i,j)* entry denotes whether nodes *i* and *j* are adjacent (in which case its value is 1) or not (value 0).

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Orthology
The evolutionary relationship between two genes in two species that have descended from a common ancestor.

(that is, to send fluid only along outgoing edges)[34], the associated algorithmics and the utility of the propagation process in this setting are still active areas of research.

Application to protein function prediction

Network propagation has a decade-long history in biology, and among its earliest applications were techniques for detecting homology between protein sequences. For example, the Rankprop algorithm[35] applied network propagation to predict protein-fold classes on a multi-species network based on protein sequence similarity; importantly, it was shown to outperform the yardstick PSI-BLAST algorithm, a profile-based ranking approach. Rankprop was later extended to account for amino acid sequence motifs[36] and PPI information[37]. Similar ideas were used by Isorank[38] and its follow-up, IsoRankN[39], to predict functional orthology by aligning PPI networks from multiple species.

This early success of network propagation and related approaches[40–42] led to the development of numerous propagation-based algorithms for protein function prediction (TABLE 1). These algorithms can be categorized based on whether they operate on a single network (with edges of potentially multiple types) or whether they integrate propagation information from multiple networks using machine-learning techniques.

Protein function prediction using a single network. In the most common prediction scenario, a single network is used to estimate the functional similarity between proteins. The proposed approaches differ in the propagation variant and in the type of networks used. For example, Can *et al.*[43] used RWR for a fixed number of time steps to predict pathway and co-complex memberships based on a PPI network: the number of times the walk landed on a protein when it was started at a random known member of the pathway or complex was used to rank all the other proteins as potential members of that pathway or complex. Voevodski *et al.*[44] suggested a PageRank affinity measure, which takes the minimum of the two random walks between a pair of proteins (that is, from the first protein to the second protein and vice versa) as a measure of how likely they are to be members of the same protein complex. Suthram *et al.*[45] used the electrical circuit formulation of network propagation to pinpoint genes that underlie expression variation. To this end, they modelled the flow of information from a potential source gene to target genes as electric currents through a protein network.

In the previous examples, the computations were done with respect to a PPI network in which it is assumed that the shorter the path between two proteins the more likely the proteins are to share similar

Table 1 | **Software tools based on network propagation**

Tool	Goal	Type	Platform	Web site
Function prediction				
DSD[48] and capDSD[34]	Function prediction	Single network	Web server and software for download	http://dsd.cs.tufts.edu/server/ and http://dsd.cs.tufts.edu/capdsd
GeneMANIA [103]	Function prediction	Single network	Cytoscape plugin	http://apps.cytoscape.org/apps/genemania
Mashup[56]	Function prediction	Integrative	Software for download	http://mashup.csail.mit.edu/
RIDDLE[70]	Function prediction	Single network	Web server	http://www.functionalnet.org/RIDDLE/
Disease characterization				
CATAPULT[82]	Gene prioritization	Integrative	Web server and software for download	http://marcottelab.org/index.php/Catapult
Cytoscape 'diffuse' service[104]	General propagation	1D and 2D	Software for download	• http://cytoscape.org • Native in version 3.5 and greater
DADA[80]	Gene prioritization	1D	Software for download	http://compbio.case.edu/dada/
Exome Walker[72]	Gene prioritization	1D	Web server	http://compbio.charite.de/ExomeWalker
GUILD[105]	Gene prioritization	1D	Software for download	http://sbi.imim.es/web/index.php/research/software/guildsoftware
HotNet2 (REF. 30)	Module detection	2D	Software for download	http://compbio.cs.brown.edu/projects/hotnet2/
NBS[89]	Patient stratification	Integrative	Software for download	http://chianti.ucsd.edu/~mhofree/NBS/
NetQTL[79]	Gene prioritization and module detection	1D	Software for download	https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/index.cgi#netqtl
PRINCIPLE[106]	Gene prioritization and module detection	1D	Cytoscape plugin	http://www.cs.tau.ac.il/~bnet/software/PrincePlugin/
SNF[90]	Patient stratification	Integrative	Software for download	http://compbio.cs.toronto.edu/SNF/SNF/Software.html
TieDIE[91]	Module detection	Integrative	Software for download	https://sysbiowiki.soe.ucsc.edu/tiedie
ToppGene[107]	Gene prioritization	1D	Web server	https://toppgene.cchmc.org/

capDSD, confidence, augmented pathway diffusion state distance; CATAPULT, combining data across species using positive-unlabeled learning techniques; DADA, degree-aware disease gene prioritization; DSD, diffusion state distance; GeneMANIA, multiple association network integration algorithm; GUILD, genes underlying inheritance linked disorders; NBS, network-based stratification; PRINCIPLE, prioritization and complex elucidation implementation; RIDDLE, reflective diffusion and local extension;SNF, similarity network fusion; TieDIE, tied diffusion through interacting events.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

functional roles. By contrast, in genetic interaction networks, the length of a path may depend on the type of relationship between two proteins. In particular, co-complex relationships might be characterized by even-length paths, as genetic interactions were found to be particularly enriched between genes belonging to different pathways or complexes (implying that genes from the same complex are likely to have even-length paths between them)[46]. This observation led Qi *et al.*[47] to devise a parity-aware variant of network propagation that takes into account odd-length versus even-length paths to predict genetic interactions and protein co-complex memberships[47].

Beyond the direct applications of network propagation, the diffusion state distance (DSD) of Cao *et al.*[34,48] represented each node (that is, protein) by a vector recording its expected random walk distance to all other nodes, and defined a distance measure between these vectors. The authors then predicted the function of a protein based on the known functions of the proteins that were closest to it in DSD. Similarly, Compass[49] used a symmetric version of network propagation to quantify the functional similarity between proteins and to thereby predict protein function. Finally, GeneMANIA[50] used a two-stage prediction process, in which multiple sources of data are first combined into a single function-specific association network, and then a propagation-based approach is applied to the resulting network to predict gene function.

Propagating over multiple networks for protein function prediction. When multiple protein networks are available, rather than merging them into a single network, several methods have been proposed that separately propagate information on each network and then integrate the results to increase the confidence in the subsequent predictions. For example, Peng *et al.*[51] carried out random walks across three different networks (representing PPIs, functional similarity and domain co-occurrence), combining the propagated scores after each propagation step as an input to the next step. More sophisticated machine-learning methods were used by Lanckriet *et al.*[52], Lee *et al.*[53] and Tsuda *et al.*[54,55] who fed the network propagation results as an input to a protein function classifier. Recently, Cho *et al.*[56] and Wang *et al.*[57] have introduced the Diffusion Component Analysis method. Their starting point was RWR, but they then applied a dimensionality reduction technique to reduce the dimension of the propagation results in a way that well-approximates the RWR matrix. They argued that the low-dimensional representation is less sensitive to noise in the network. This scheme was also generalized to handle multiple molecular networks simultaneously, producing state-of-the-art results in predicting gene function and genetic interactions[56].

Other important work includes the use of network propagation to discover network modules, or communities, in a PPI network[58–60]. For example, Macropol *et al.*[60] ranked nodes (representing proteins) using RWR to greedily grow candidate modules; each time, the highest scoring node with respect to RWR from the members of a current module was added to the module. This process continued until the score of the node to be added fell below a pre-set threshold or until a maximum number of nodes had been reached.

Network propagation has also been used, although to a lesser extent, in many other applications. These applications include ranking differentially expressed genes[61], predicting gene essentiality and pleiotropy (that is, association with multiple phenotypes)[62,63], identifying signalling-regulatory pathways[64,65], predicting drug side effects[66], reducing noise in PPI networks[67,68], defining functional similarity[69] and predicting functional associations of unannotated gene sets[70]. This diverse set of applications underscores the generality and utility of the propagation technique.

Application to human disease

Capitalizing on the multiple successful biological applications, network propagation techniques are now being applied to the study of human diseases (TABLE 1). These applications can be classified into three broad categories (FIG. 3): 1D methods that seek to score genes; 2D methods that score gene–gene similarities and that use these scores to derive gene modules; and integrative approaches that combine multiple 1D computations using multiple data sources.

Scoring genes with 1D network propagation methods. The first applications of network propagation to study human disease were for gene prioritization. Multiple researchers[71–77] aimed to predict causal genes for a disease by starting from known causal genes of similar diseases (known as 'seeds') and applying network propagation to prioritize genes based on their proximity to the prior seed set. Although the techniques varied slightly in the way the propagation or diffusion was carried out, the same propagation engine was common to all. For example, Nitsch *et al.*[78] suggested different random walk models for gene prioritization based on differential gene expression data, and found that the standard network propagation technique (which they named heat kernel diffusion ranking) performs best[78]. In another example, Kim *et al.*[79] used an electrical circuit model to find physical pathways that connect copy number variations of potentially causal genes to mRNA expression changes of target genes between disease and control cases. Applying their approach to glioblastoma data, they were able to uncover candidate causal genes and pathways that potentially explained the expression changes.

The disease gene prioritization framework was later extended in several ways. The DADA approach[80] corrected for bias in prioritization scores that was due to network topology (that is, the overall arrangement of the nodes and edges in the network) by estimating the significance of each score while explicitly accounting for node degree. Erten *et al.*[81] applied a 2D method for gene prioritization, in which each gene was scored by its similarity to every other gene in a PPI network based on network propagation. These scores were then used to rank candidate genes by their similarity to known disease genes. The main novelty in this work compared with previous 1D approaches

Classifier
A machine-learning algorithm that predicts the class of a sample given some characteristics of it. For example, a classifier can aim to distinguish between disease and non-disease genes based on their network proximity to known disease or non-disease genes.

Network modules
Regions of a network with some topological property; for example, a set of nodes that densely interact with one another.

Node degree
The number of other nodes that are adjacent (that is, directly connected) to a node.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Similarity matrix
A matrix with rows and columns that represent the same set of objects such that the *(i,j)* entry denotes some similarity measure (for example, as obtained from network propagation) between the corresponding elements.

is that candidate genes are evaluated based on their topological similarity to disease genes rather than on their proximity. Singh-Blom *et al*[82]. developed the CATAPULT (combining data across species using positive-unlabeled learning techniques) algorithm[82] that applies a 'truncated' version of network propagation (considering walk lengths of up to six steps) on a hybrid network that contains both genes and phenotypes; the authors argue that longer paths are less informative than shorter paths, with the contribution of paths to similarity scoring becoming less substantial with increasing path length. The resulting scores of paths of different lengths and edge types were rigorously combined by feeding them as features to a classifier of gene–phenotype associations.

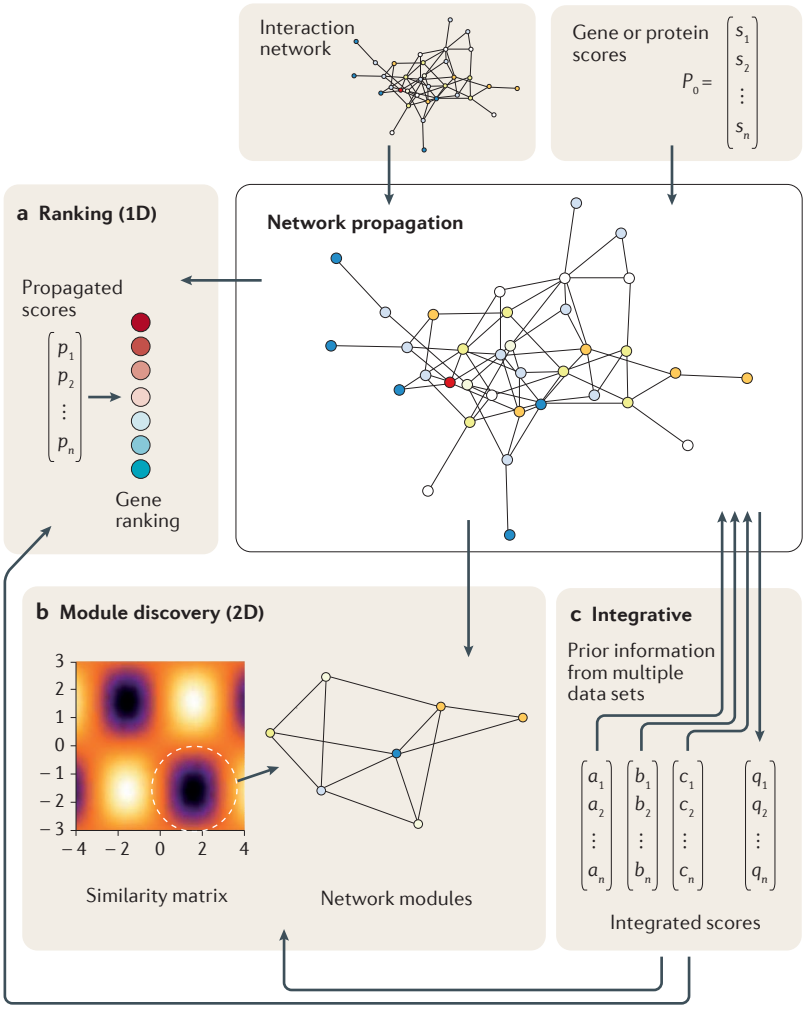

Figure 3 | **Overview of approaches that use network propagation.** Network propagation approaches take a vector the entries of which (0 or 1 or real-values) indicate the prior information on each gene or node in the network. Following propagation, the scores on the nodes are examined using different approaches. **a** | 1D approaches rank or prioritize genes by their propagated scores. **b** | 2D approaches analyse a similarity matrix defined by the propagation and extract modules, or subnetworks, according to both the propagated scores and the topology of the network. **c** | Integrative approaches propagate prior information from different data sets, or individuals, across one or more networks, forming integrated scores that are used to rank genes and/or to extract modules.

Deriving gene modules using 1D and 2D network propagation methods. Complex diseases are thought to be caused by disease modules, which comprise multiple genes that function together to drive specific cellular processes. Network propagation methods were shown to be instrumental in the inference of disease modules[83] and have been used in two different modes. The first (1D) mode operates by projecting the gene-based scores on a functional, or physical, interaction network, seeking dense regions of this network that span high-scoring genes[74,84,85]. The second (2D) mode instead operates on the similarity matrix that records the propagation value of every protein when using every other protein as a single prior (BOX 1). This similarity matrix is clustered in various ways to reveal disease modules. This approach was used by the HotNet[86] and HotNet2 (REF. 30) algorithms to find modules of somatic mutations in cancer and modules of common variants in complex diseases[87]. HotNet and HotNet2 differ in the diffusion processes and clustering algorithms that they use, with HotNet2 accounting for the asymmetry in the similarity matrix that results from propagation. FIGURE 4a demonstrates the power of this approach in detecting modules whose individual genes are not necessarily highly mutated but co-occur in close proximity on the propagated network.

Combining multiple data sets with integrative network propagation methods. In addition to the 1D gene prioritization and 2D module discovery approaches, integrative approaches that combine multiple 1D computations have emerged. These methods integrate multiple lines of evidence to improve prediction performance.

For example, Rufallo *et al.*[31] have shown that integrative network propagation is a powerful tool for pinpointing cancer driver genes (FIG. 2c). In this study, propagation was used to integrate somatic mutation and differential expression data from cancer patients, creating combined features that were fed to a classifier for the prediction of causal genes in specific cancer types. In a similar study, HIT'n DRIVE[88] prioritized cancer driver genes by connecting somatic mutations or copy-number aberrations to downstream expression changes; in this case, a different mathematical formulation was used based on the hitting time of a random walk, which was defined as the expected number of hops that a random walk from a source node takes to first reach a target node.

Network propagations can also serve as a basis for stratifying patients to identify disease subtypes[89] (FIG. 4b). The underlying principle is to use network propagation to combat heterogeneity in genetic variants or mutations across a patient population. Although patients may have variants that affect very different sets of genes, network propagation of the data from each patient highlights similar network regions, allowing these patients to be clustered together (FIG. 1c). The application of this approach to uterine, ovarian and lung cancer data led to the identification of putative subtypes and the association of disease modules with these subtypes. A second example of the use of propagation to integrate multiple information sources for patient stratification is

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

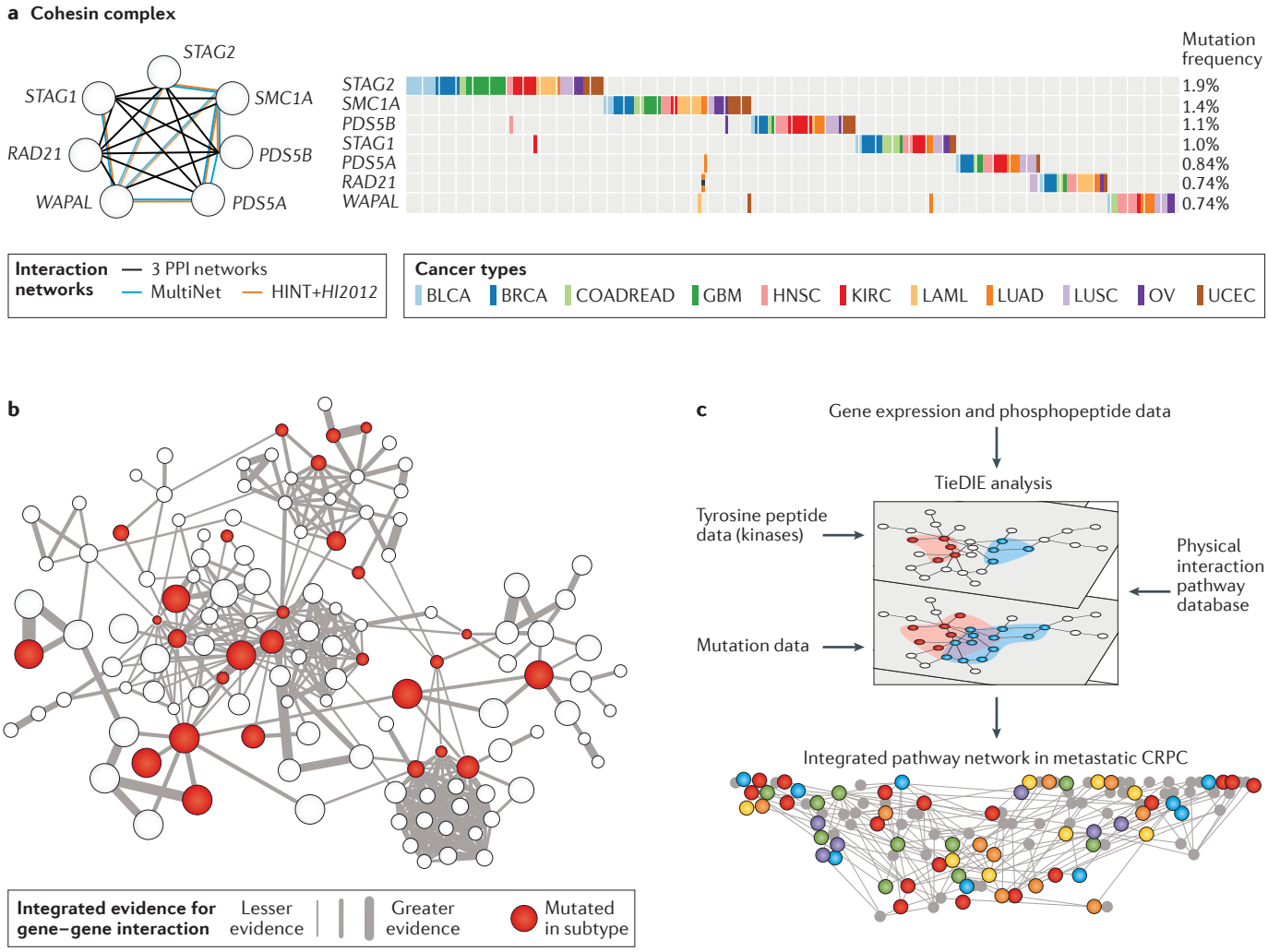

Figure 4 | **Applications of network propagation to analyse cancer data.** **a** | Identifying disease modules with network propagation. The cohesin protein complex is identified using HotNet2, a 2D approach, by propagating the frequencies of somatic mutations in 12 cancer types from The Cancer Genome Atlas. The right panel shows a mutation matrix, the rows of which are the genes in the identified module, and the columns of which are the samples (colour-coded by cancer type) that have a mutation in these genes. Each of the genes in the module is mutated at extremely low frequency, but the propagation of individual frequencies across the network amplifies this weak signal, as these genes are connected by many edges across multiple protein–protein interaction (PPI) networks (left panel). **b** | Patient stratification with network propagation. The network-based stratification (NBS) integrative approach is used to identify a robust cluster of patients with ovarian cancer, suggesting a new disease subtype. A subnetwork of genes that have high propagated mutation scores in this patient cluster (denoted by node size) and that is most responsible for discriminating the somatic mutation profiles of this subtype from others, is shown. Edge width reflects confidence. Filled nodes indicate that somatic mutations were found for the corresponding gene in the examined cohort. **c** | The TieDIE (tied diffusion through interacting events) integrative approach is used to integrate gene expression, somatic mutations and phosphoproteomic data in castration-resistant prostate cancer (CRPC) to link genomic mutations, kinase regulators and transcription regulators. A 'scaffold' network that was generated by TieDIE and is centred on six cancer hallmark categories is shown. Hallmark genes are colour-coded according to their annotated category. Other network genes that connect two or more of these hallmark genes are shown in grey. BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; COAD, colon adenocarcinoma; READ, rectum adenocarcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KIRC, kidney renal clear cell carcinoma; LAML, acute myeloid leukemia; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; UCEC, uterine corpus endometrioid carcinoma. Note that data for COAD and READ have been combined. Part **a** is adapted with permission from REF. 30, Macmillan Publishers Limited. Part **b** is adapted with permission from REF. 89, Macmillan Publishers Limited. Part **c** is adapted with permission from REF. 92, Elsevier.

the similarity network fusion approach[90]. In this work, Wang *et al.* performed propagation across patient similarity networks that were derived from individual data types (for example, mutation, mRNA expression and DNA methylation data) to form a single patient similarity network, which was used to derive patient subtypes with distinct survival profiles.

Another application of network propagation methods is for the identification of disease modules. For example, the TieDIE (tied diffusion through interacting events) approach[91] searches for disease modules in cancer by performing two propagation computations: one starting from mutated genes and another starting from differentially expressed genes.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

By combining the resulting rankings (taking the minimum and thresholding the result), Paull *et al.* could search for a subnetwork that connected the high-scoring genes. In a follow-up paper[92], TieDIE was used in combination with phosphoproteomic data to identify active pathways in prostate cancer and to derive patient-specific network models and potential treatment strategies (FIG. 4c).

Network propagation methods can also be applied to facilitate the prediction of novel drug targets. For example, Shnaps *et al.*[93] used network propagation to simulate the effect of a drug, which targeted a candidate protein, on patients with acute myeloid leukaemia. To this end, they executed network propagation for one patient at a time, using either the complete PPI network or the network from which the candidate protein was removed. Focusing on the differentially expressed genes (tumour versus normal) of each patient, Shnaps *et al.* ranked each candidate protein according to the change that its removal induced on the propagated values of these genes. In a different study, rather than simulating the effect of a drug, Chen *et al.*[94] aimed to predict drug–target relationships based on the assumption that similar drugs target similar proteins. To this end, they successfully applied network propagation to predict drug–target relationships by using an integrated network that included target–target, drug–target and drug–drug relationships.

Overall, network propagation techniques are becoming increasingly abundant and are producing state-of-the-art results for a wide variety of applications in disease genomics, ranging from gene prioritization to genetic module identification and drug target prediction.

Perspectives and conclusions

In this Review, we have described the method of network propagation, emphasizing the generality and power of the approach. We have reviewed some of the various applications of network propagation in biology, focusing on gene ranking (1D), module identification (2D) and integrative applications of network propagation for the study of human disease. In gene ranking, propagation methods were shown to amplify weak associations of genes with phenotypes. In module identification, network propagation methods allowed the inclusion in modules of genes for which little or no direct evidence of involvement was available. In integrative applications, network approaches amplified weak similarities between different sources of information (such as different molecular species or different patients (FIG. 1c)) to create robust similarities that can be used to build patient similarity and disease similarity networks. We conclude that network propagation is an essential tool in any genetic toolbox that seeks to leverage network information in the study of the genes and genetic modules that drive disease.

Despite its general applicability and good performance, the basic network propagation scheme might be improved in several ways to allow more rigorous scoring. In particular, the scheme assumes that the contribution of a path to the propagation score diminishes exponentially with its length. This assumption can be removed to allow different weights for different path lengths. For example, the CATAPULT algorithm[82] learns a weighting in a supervised manner, which also accounts for the types of edges included in the propagation paths.

Another important scoring issue is the evaluation of the significance of a given propagation score. Several authors have replaced the propagation scores with *P* values to resolve this concern. This approach has the desired effect of down-weighting hubs, which tend to have high propagation scores. For example, Mazza *et al.*[85] computed an empirical *P* value for every node by observing the distribution of scores of that particular node under the propagations of randomized prior vectors.

We have presented two main ways in which network propagation can be applied to a single network: in a 1D setting in which the output is the final scoring vector p, and in a 2D mode in which the output is the corresponding transformation, or similarity matrix. There are also several ways in which one could use the prior information (for example, binary versus continuous). These different choices may affect the analysis, and the best performing variant will be application-specific.

Although most studies have so far focused on a single network, the integration of multiple networks has been repeatedly shown to improve the predictive power of different methods[22]. This observation reinforces the value of combining multiple data sources — for example, using tissue-specific, cell type-specific and/or patient-specific networks — to improve the downstream analysis. There has been promising research into developing such networks[84,95], with important developments underway from the Genotype-Tissue Expression (GTEx) project, which enables the construction of a specific network for each major human tissue[96]. In addition, regulatory networks are increasing in scope and resolution, through efforts such as ENCODE[97] and the Roadmap Epigenomics project[98]. Thus, network propagation will continue to be a powerful method for integrating an increasingly diverse collection of scores across a wide range of biological interaction networks, leading to deeper insights into biological processes and disease.

1. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
2. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
3. Schwikowski, B., Uetz, P. & Fields, S. A network of protein–protein interactions in yeast. *Nat. Biotechnol.* **18**, 1257–1261 (2000).
4. Brohée, S. & van Helden, J. Evaluation of clustering algorithms for protein–protein interaction networks. *BMC Bioinformatics* **7**, 488 (2006).
5. Song, J. & Singh, M. How and when should interactome-derived clusters be used to predict functional modules and protein function? *Bioinformatics* **25**, 3143–3150 (2009).
6. Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Mol. Syst. Biol.* **3**, 88 (2007).
7. Peña-Castillo, L. *et al.* A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence. *Genome Biol.* **9** (Suppl. 1), S2 (2008).
8. Navlakha, S. & Kingsford, C. The power of protein interaction networks for associating genes with diseases. *Bioinformatics* **26**, 1057–1063 (2010).
9. Menche, J. *et al.* Uncovering disease–disease relationships through the incomplete interactome. *Science* **347**, 1257601–1257601 (2015).

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

10. Shrager, J., Hogg, T. & Huberman, B. A. Observation of phase transitions in spreading activation networks. *Science* **236**, 1092–1094 (1987).
11. Lovász, L. in *Combinatorics: Paul Erdős is Eighty* (eds Miklós, D., Sós, V. T. & Szőnyi, T.), 1–46 (Janos Bolyai Mathematical Society, 1993.
12. Page, L., Brin, S., Motwani, R. & Winograd, T. The PageRank citation ranking: bringing order to the web. *Stanford InfoLab* http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.1768 (1999).
13. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *J. of the ACM* **46**, 604–632 (1999).
14. Klein, D. J. & Randic´, M. Resistance distance. *J. Math. Chem.* **12**, 81–95 (1993).
15. Tong, H., Faloutsos, C. & Pan, J.-Y. Random walk with restart: fast solutions and applications. *Knowl. Inf. Syst.* **14**, 327–346 (2007).
16. Haveliwala, T. H. Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search. *IEEE Trans. Knowl. Data Eng.* **15**, 784–796 (2003).
17. Krapivsky, P. L., Redner, S. & Ben-Naim, E. *A Kinetic View of Statistical Physics* (Cambridge Univ. Press, 2010).
18. Ben-Avraham, D. & Havlin, S. *Diffusion and Reactions in Fractals and Disordered Systems* (Cambridge Univ. Press, 2000).
19. Doyle, P. G. & Laurie Snell, J. *Random Walks and Electric Networks* (The Mathematical Association of America, 1984).
20. Kondor, R. I. & Lafferty, J. Diffusion kernels on graphs and other discrete input spaces. *Proc. Intl Conf. on Machine Learning (ICML)* **2**, 315–322 (2002).
21. Noble, W. S., Kuang, R., Leslie, C. & Weston, J. Identifying remote protein homologs by network propagation. *FEBS J.* **272**, 5119–5128 (2005).
22. Mitra, K., Carvunis, A.-R., Ramesh, S. K. & Ideker, T. Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* **14**, 719–732 (2013).
23. Cho, D.-Y., Kim, Y.-A. & Przytycka, T. M. Chapter 5: network biology approach to complex diseases. *PLoS Comput. Biol.* **8**, e1002820 (2012).
24. Ideker, T. & Sharan, R. Protein networks in disease. *Genome Res.* **18**, 644–652 (2008).
25. Csermely, P., Korcsmáros, T., Kiss, H. J. M., London, G. & Nussinov, R. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Ther.* **138**, 333–408 (2013).
26. Oti, M., Snel, B., Huynen, M. A. & Brunner, H. G. Predicting disease genes using protein–protein interactions. *J. Med. Genet.* **43**, 691–698 (2006).
27. Franke, L. *et al.* Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.* **78**, 1011–1025 (2006).
28. Barabasi, A.-L. Scale-free networks: a decade and beyond. *Science* **325**, 412–413 (2009).
29. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
30. Leiserson, M. D. M. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).
A 2D method that exploits the propagation-derived similarity matrix to infer protein modules that are associated with cancer.
31. Ruffalo, M., Koyutürk, M. & Sharan, R. Network-based integration of disparate omic data to identify 'silent players' in cancer. *PLoS Comput. Biol.* **11**, e1004595 (2015).
32. Du, D., Lee, C. F. & Li, X.-Q. Systematic differences in signal emitting and receiving revealed by PageRank analysis of a human protein interactome. *PLoS ONE* **7**, e44872 (2012).
33. Vinayagam, A. *et al.* A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signal.* **4**, rs8 (2011).
34. Cao, M. *et al.* New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics* **30**, i219–i227 (2014).
A network propagation-based approach for incorporating known biological pathways into protein function prediction.
35. Weston, J., Elisseeff, A., Zhou, D., Leslie, C. S. & Noble, W. S. Protein ranking: from local to global structure in the protein similarity network. *Proc. Natl Acad. Sci. USA* **101**, 6559–6563 (2004).
One of the first studies to apply the concept of network propagation to the biological domain.
A propagation process over sequence similarity networks of different species is used to predict orthology.
36. Kuang, R., Weston, J., Noble, W. S. & Leslie, C. Motif-based protein ranking by network propagation. *Bioinformatics* **21**, 3711–3718 (2005).
37. Yosef, N., Sharan, R. & Noble, W. S. Improved network-based identification of protein orthologs. *Bioinformatics* **24**, i200–i206 (2008).
38. Singh, R., Xu, J. & Berger, B. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc. Natl Acad. Sci. USA* **105**, 12763–12768 (2008).
39. Liao, C.-S., Lu, K., Baym, M., Singh, R. & Berger, B. IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics* **25**, i253–i258 (2009).
40. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B. & Singh, M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* **21** (Suppl. 1), i302–i310 (2005).
41. Letovsky, S. & Kasif, S. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics* **19** (Suppl. 1), i197–i204 (2003).
42. Deng, M., Zhang, K., Mehta, S., Chen, T. & Sun, F. Prediction of protein function using protein–protein interaction data. *J. Comput. Biol.* **10**, 947–960 (2003).
43. Can, T., Çamoglu, O. & Singh, A. K. Analysis of protein–protein interaction networks using random walks. *BIOKDD '05* https://doi.org/10.1145/1134030.1134042 (2005).
44. Voevodski, K., Teng, S.-H. & Xia, Y. Spectral affinity in protein networks. *BMC Syst. Biol.* **3**, 112 (2009).
45. Suthram, S., Beyer, A., Karp, R. M., Eldar, Y. & Ideker, T. eQED: an efficient method for interpreting eQTL associations using protein networks. *Mol. Syst. Biol.* **4**, 162 (2008).
46. Kelley, R. & Ideker, T. Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.* **23**, 561–566 (2005).
47. Qi, Y., Suhail, Y., Lin, Y.-Y., Boeke, J. D. & Bader, J. S. Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Res.* **18**, 1991–2004 (2008).
48. Cao, M. *et al.* Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PLoS ONE* **8**, e76339 (2013).
49. Lehtinen, S., Lees, J., Bähler, J., Shawe-Taylor, J. & Orengo, C. Gene function prediction from functional association networks using kernel partial least squares regression. *PLoS ONE* **10**, e0134668 (2015).
50. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.* **9** (Suppl. 1), S4 (2008).
51. Peng, W., Li, M., Chen, L. & Wang, L. Predicting protein functions by using unbalanced random walk algorithm on three biological networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **14**, 360–369 (2015).
52. Lanckriet, G. R. G., De Bie, T., Cristianini, N., Jordan, M. I. & Noble, W. S. A statistical framework for genomic data fusion. *Bioinformatics* **20**, 2626–2635 (2004).
53. Lee, H., Tu, Z., Deng, M., Sun, F. & Chen, T. Diffusion kernel-based logistic regression models for protein function prediction. *OMICS* **10**, 40–55 (2006).
54. Tsuda, K., Shin, H. & Schölkopf, B. Fast protein classification with multiple networks. *Bioinformatics* **21** (Suppl. 2), ii59–ii65 (2005).
55. Tsuda, K. & Noble, W. S. Learning kernels from biological networks by maximizing entropy. *Bioinformatics* **20** (Suppl. 1), i326–i333 (2004).
56. Cho, H., Berger, B. & Peng, J. Compact integration of multi-network topology for functional analysis of genes. *Cell Syst.* **3**, 540–548.e5 (2016).
An integrative network propagation approach for functional inference using multiple heterogeneous networks.
57. Wang, S., Cho, H., Zhai, C., Berger, B. & Peng, J. Exploiting ontology graph for predicting sparsely annotated gene function. *Bioinformatics* **31**, i357–i364 (2015).
58. Voevodski, K., Teng, S.-H. & Xia, Y. Finding local communities in protein networks. *BMC Bioinformatics* **10**, 297 (2009).
59. Peng, W., Wang, J., Zhao, B. & Wang, L. Identification of protein complexes using weighted PageRank-nibble algorithm and core-attachment structure. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **12**, 179–192 (2015).
60. Macropol, K., Can, T. & Singh, A. K. RRW: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinformatics* **10**, 283 (2009).
61. Morrison, J. L., Breitling, R., Higham, D. J. & Gilbert, D. R. GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics* **6**, 233 (2005).
62. Missiuro, P. V. *et al.* Information flow analysis of interactome networks. *PLoS Comput. Biol.* **5**, e1000350 (2009).
63. Zotenko, E., Mestre, J., O'Leary, D. P. & Przytycka, T. M. 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 (2008).
64. Tu, Z., Wang, L., Arbeitman, M. N., Chen, T. & Sun, F. An integrative approach for causal gene identification and gene regulatory pathway inference. *Bioinformatics* **22**, e489–e496 (2006).
65. Yeger-Lotem, E. *et al.* Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat. Genet.* **41**, 316–323 (2009).
66. Atias, N. & Sharan, R. An algorithmic framework for predicting side effects of drugs. *J. Comput. Biol.* **18**, 207–218 (2011).
67. Lei, C. & Ruan, J. A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity. *Bioinformatics* **29**, 355–364 (2013).
68. Alkan, F. & Erten, C. RedNemo: topology-based PPI network reconstruction via repeated diffusion with neighborhood modifications. *Bioinformatics* **33**, 537–544 (2016).
69. Lerman, G. & Shakhnovich, B. E. Defining functional distance using manifold embeddings of gene ontology annotations. *Proc. Natl Acad. Sci. USA* **104**, 11334–11339 (2007).
70. Wang, P. I. *et al.* RIDDLE: reflective diffusion and local extension reveal functional associations for unannotated gene sets via proximity in a gene network. *Genome Biol.* **13**, R125 (2012).
71. Li, Y. & Patra, J. C. Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. *Bioinformatics* **26**, 1219–1224 (2010).
72. Smedley, D. *et al.* Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics* **30**, 3215–3222 (2014).
73. Köhler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* **82**, 949–958 (2008).
An application of network propagation to prioritize disease-causing genes.
74. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. & Sharan, R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.* **6**, e1000641 (2010).
One of the first studies to use network propagation to associate modules of multiple proteins with disease.
75. Lee, I., Blom, U. M., Wang, P. I., Shim, J. E. & Marcotte, E. M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* **21**, 1109–1121 (2011).
76. Chen, J., Aronow, B. J. & Jegga, A. G. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics* **10**, 73 (2009).
77. Chen, J. Y., Shen, C. & Sivachenko, A. Y. Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Pac. Symp. Biocomput.* **2006**, 367–378 (2006).
78. Nitsch, D., Gonçalves, J. P., Ojeda, F., de Moor, B. & Moreau, Y. Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics* **11**, 460 (2010).
79. Kim, Y.-A., Wuchty, S. & Przytycka, T. M. Identifying causal genes and dysregulated pathways in complex diseases. *PLoS Comput. Biol.* **7**, e1001095 (2011).
80. Erten, S., Bebek, G., Ewing, R. M. & Koyutürk, M. DADA: degree-aware algorithms for network-based disease gene prioritization. *BioData Min.* **4**, 19 (2011).

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

81. Erten, S., Bebek, G. & Koyutürk, M. Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. *J. Comput. Biol.* **18**, 1561–1574 (2011).
82. Singh-Blom, U. M. *et al.* Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PLoS ONE* **8**, e58977 (2013).
83. Kim, Y.-A., Cho, D.-Y. & Przytycka, T. M. Understanding genotype–phenotype effects in cancer via network approaches. *PLoS Comput. Biol.* **12**, e1004747 (2016).
84. Magger, O., Waldman, Y. Y., Ruppin, E. & Sharan, R. Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Comput. Biol.* **8**, e1002690 (2012).
85. Mazza, A., Klockmeier, K., Wanker, E. & Sharan, R. An integer programming framework for inferring disease complexes from network data. *Bioinformatics* **32**, i271–i277 (2016).
86. Vandin, F., Upfal, E. & Raphael, B. J. Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* **18**, 507–522 (2011).
87. Nakka, P., Raphael, B. J. & Ramachandran, S. Gene and network analysis of common variants reveals novel associations in multiple complex diseases. *Genetics* **204**, 783–798 (2016).
88. Shrestha, R. *et al.* in *Research in Computational Molecular Biology. RECOMB 2014. Lecture Notes in Computer Science* (ed. Sharan, R.) 293–306 (Springer, 2014).
89. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **10**, 1108–1115 (2013). **One of the first methods to use patient-specific propagation processes to stratify patients with cancer into subtypes.**
90. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337 (2014).
91. Paull, E. O. *et al.* Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics* **29**, 2757–2764 (2013). **An integrative method to predict cancer pathways that is based on superimposing two propagation processes that are run from nodes corresponding to mutated and differentially expressed genes.**
92. Drake, J. M. *et al.* Phosphoproteome integration reveals patient-specific networks in prostate cancer. *Cell* **166**, 1041–1054 (2016).
93. Shnaps, O., Perry, E., Silverbush, D. & Sharan, R. Inference of personalized drug targets via network propagation. *Pac. Symp. Biocomput.* **21**, 156–167 (2016).
94. Chen, X., Xing, C., Ming-Xi, L. & Gui-Ying, Y. Drug–target interaction prediction by random walk on the heterogeneous network. *Mol. Biosyst.* **8**, 1970 (2012).
95. Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47**, 569–576 (2015).
96. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
97. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc. Natl Acad. Sci. USA* **111**, 6131–6138 (2014).
98. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
99. Chung, F. Laplacians and the Cheeger inequality for directed graphs. *Ann. Comb.* **9**, 1–19 (2005).
100. Malliaros, F. D. & Vazirgiannis, M. Clustering and community detection in directed networks: a survey. *Phys. Rep.* **533**, 95–142 (2013).
101. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
102. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).
103. Montojo, J. *et al.* GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics* **26**, 2927–2928 (2010).
104. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
105. Guney, E. & Oliva, B. Exploiting protein–protein interaction networks for genome-wide disease-gene prioritization. *PLoS ONE* **7**, e43557 (2012).
106. Gottlieb, A., Magger, O., Berman, I., Ruppin, E. & Sharan, R. PRINCIPLE: a tool for associating genes with diseases via network propagation. *Bioinformatics* **27**, 3325–3326 (2011).
107. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, W305–W311 (2009).

Acknowledgements
The authors gratefully acknowledge J. Huang and M. Ruffalo for assistance with figures for this manuscript. They also thank E. Eisenberg for assistance with references for this manuscript. This work was initiated while the authors attended a Network Biology workshop as part of a semester on Algorithmic Challenges in Genomics at the Simons Institute for the Theory of Computing at University of California, Berkeley, USA.

Competing interests statement
The authors declare competing interests: see Web version for details.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.