Bioinformatics, 35(3), 2019, 497–505
doi: 10.1093/bioinformatics/bty637
Advance Access Publication Date: 18 July 2018
Original Paper

Systems biology

Random walk with restart on multiplex and heterogeneous biological networks

Alberto Valdeolivas[1,2,*], Laurent Tichit[1], Claire Navarro[2,3], Sophie Perrin[2,3], Gaëlle Odelin[2,3], Nicolas Levy[3], Pierre Cau[2,3], Elisabeth Remy[1] and Anaïs Baudot[1,*]

[1]Aix Marseille Univ, CNRS, Centrale Marseille, I2M, 13009, Marseille, France, [2]ProGeLife, 13001, Marseille and [3]Aix Marseille Univ, INSERM, MMG, 13005, Marseille, France

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on September 20, 2017; revised on June 13, 2018; editorial decision on July 11, 2018; accepted on July 16, 2018

Abstract

Motivation: Recent years have witnessed an exponential growth in the number of identified interactions between biological molecules. These interactions are usually represented as large and complex networks, calling for the development of appropriated tools to exploit the functional information they contain. Random walk with restart (RWR) is the state-of-the-art guilt-by-association approach. It explores the network vicinity of gene/protein seeds to study their functions, based on the premise that nodes related to similar functions tend to lie close to each other in the networks.

Results: In this study, we extended the RWR algorithm to multiplex and heterogeneous networks. The walk can now explore different layers of physical and functional interactions between genes and proteins, such as protein–protein interactions and co-expression associations. In addition, the walk can also jump to a network containing different sets of edges and nodes, such as phenotype similarities between diseases. We devised a leave-one-out cross-validation strategy to evaluate the algorithms abilities to predict disease-associated genes. We demonstrate the increased performances of the multiplex-heterogeneous RWR as compared to several random walks on monoplex or heterogeneous networks. Overall, our framework is able to leverage the different interaction sources to outperform current approaches. Finally, we applied the algorithm to predict candidate genes for the Wiedemann–Rautenstrauch syndrome, and to explore the network vicinity of the SHORT syndrome.

Availability and implementation: The source code is available on GitHub at: https://github.com/alberto-valdeolivas/RWR-MH. In addition, an R package is freely available through Bioconductor at: http://bioconductor.org/packages/RandomWalkRestartMH/.

Contact: alberto.valdeolivas@etu.univ-amu.fr or anais.baudot@univ-amu.fr

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

1 Introduction

Recent years have witnessed the accumulation of physical and functional interactions between biological macromolecules. For instance, protein–protein interactions (PPI) are nowadays screened at the proteome scale for many organisms, revealing thousands of physical interactions between proteins. Interaction data are commonly represented as networks, in which the nodes correspond to genes or proteins, and the edges to their interactions. The availability of large-scale PPI networks led to the application of graph-theory based approaches for their exploration, with the ultimate goal of extracting the knowledge they contain about cellular functioning. These methods exploit the tendency of functionally-related proteins

to lie in the same network neighborhood. For instance, clustering algorithms allow identifying communities of proteins involved in the same biological processes (Arroyo *et al.*, 2015; Brohée and van Helden, 2006; Chapple *et al.*, 2015; Katsogiannou *et al.*, 2014), and guilt-by-association strategies explore topological relationships to predict protein cellular functions (Schwikowski *et al.*, 2000).

Network-based guilt-by-association strategies, in particular, have been widely used to identify new disease-associated genes. The first approaches were parsing the direct interactors of disease proteins in a PPI network (Oti *et al.*, 2006). Then, more elaborated algorithms computing the shortest paths between candidates and known disease proteins were developed (Franke *et al.*, 2006; George *et al.*, 2006). But algorithms able to exploit the global topology, such as network propagation or random walk, were finally shown to largely outperform initial methods in the identification of disease genes (Köhler *et al.*, 2008; Vanunu *et al.*, 2010).

Random walks were first developed to explore the global topology of networks, by simulating a particle that iteratively moves from a node to a randomly selected neighboring node (Lovász, 1993). The idea of restart, which led to the random walk with restart (RWR) algorithm, was first introduced for Internet search engines. It intent to simulate the behavior of an internet user. The user surfs randomly from a web page to another thanks to the hyper-links, but he can also restart the navigation in a new arbitrary web page. Thereby, depending on the topological structure of the pages and hyper-links, some pages will be visited more frequently than others. The number of visits is considered as a proxy measure of each web page relevance (Brin and Page, 1998). Moreover, if one forces the particle to always restart in the same node or set of nodes—called seed(s)—RWR can be used to measure a proximity between the seed(s) and all the other nodes in the network (Pan *et al.*, 2004).

RWR became the state-of-the-art guilt-by-association algorithm in network computational biology. It was initially applied, as commented above, to prioritize candidate disease genes. All the network nodes are ranked by the RWR algorithm according to their proximity to known disease-associated nodes taken as seeds (Köhler *et al.*, 2008). Several extensions of the RWR algorithm further improved the prediction of disease candidate genes, mainly by considering also phenotype data (Li and Li, 2012; Li and Patra, 2010; Xie *et al.*, 2015; Zhao *et al.*, 2015). For instance, Li and Patra (2010) described a RWR on a heterogeneous network. A heterogeneous network is composed of two networks, each having its own nodes and edges, which belong to different categories, and which are linked through bipartite interactions. Li and Patra (2010) connected a PPI network with a disease–disease similarity network using known bipartite gene-disease associations.

However, a common feature and limitation of these approaches is that they perform the walks in a single network of interactions between genes and proteins. Doing so, they ignore a rich variety of information on physical and functional relationships between biological macromolecules. Indeed, not only PPI are nowadays described on a large-scale: immuno-precipitation experiments followed by mass-spectrometry can inform on the *in vivo* molecular complexes (Ruepp *et al.*, 2010), pathways interaction data are cured and stored in dedicated databases such as Reactome (Fabregat *et al.*, 2016) and Kegg (Kanehisa *et al.*, 2008). In addition, other functional interactions can be derived, for instance from transcriptomics expression data by constructing a co-expression network, or from gene ontology (GO) annotations (Ashburner *et al.*, 2000) by constructing a co-annotation network.

Each interaction source has its own meaning, relevance and bias: some networks contain links of high relevance (e.g. curated signaling pathways), while others contain thousands or even millions of interactions prone to noise (e.g. co-expression networks) (Didier *et al.*, 2015). The combination of the different sources is expected to provide a complementary view on gene and protein cellular functioning (Menche *et al.*, 2015). But networks can be combined in different ways. Generally, the different networks are merged into an aggregated network. For instance, Li and Li (2012) adapted the RWR algorithm to a network in which PPI and co-annotation interactions were aggregated. However, aggregating interaction sources as a single network dismisses the individual features and topologies of each network. In this context, the multiplex framework offers an interesting alternative. Collections of networks sharing the same nodes, but in which the edges belong to different categories or represent interactions of a different nature are called multiplex (alt. multi-slice, multi-layer) networks (Battiston *et al.*, 2014). In a biological multiplex network, each layer contains a different category of physical and functional interactions between genes or proteins.

We present here two extensions of the RWR algorithm to explore multiplex networks (RWR-M) and multiplex-heterogeneous networks (RWR-MH). We constructed a multiplex network composed of three layers of physical and functional interactions between genes and proteins, and a disease–disease network based on phenotype similarities. We applied a leave-one-out cross-validation (LOOCV) strategy to compare the RWR-M and RWR-MH algorithms to alternatives, including RWR on monoplex networks, aggregated networks and heterogeneous-only networks. We showed that considering many interaction sources through a multiplex-heterogeneous network framework enhances remarkably the performances of disease-gene prioritization. Finally, we applied the RWR-MH algorithm to predict candidate genes for being implicated in the Wiedemann–Rautenstrauch syndrome (WRS), whose responsible gene(s) remain unknown. We also explored the network vicinity of the SHORT syndrome (SS) and its associated gene, *PIK3R1*, and unveiled associated syndromes and pathways.

2 Materials and methods

2.1 Random walk on graphs

Let us consider an undirected graph, $G = (V, E)$ with adjacency matrix A. An imaginary particle starts a random walk at an initial node $v_0 \in V$. Considering the time is discrete, $t \in \mathbb{N}$, at the t-th step the particle is at node v_t. Then, it walks from v_t to v_{t+1}, a randomly selected neighbor of v_t following matrix M (Lovász, 1993). Therefore, we can write: $\forall x, y \in V, \forall t \in \mathbb{N}$

$$\mathbb{P}(v_{t+1} = y \mid v_t = x) = \begin{cases} \frac{1}{d(x)} & \text{if } (x, y) \in E \\ 0 & \text{otherwise,} \end{cases}$$

where $d(x)$ is the degree of x in the graph G. Defining $p_t(v)$ as the probability for the random walk to be at node v at time t, we can describe the evolution of the probability distribution, $\mathbf{p}_t = (p_t(v))_{v \in V}$, with the equation:

$$\mathbf{p}_{t+1}^T = M\mathbf{p}_t^T \quad (1)$$

where M denotes a transition matrix that is the column normalization of A. The stationary distribution, solution of the equation $\mathbf{p}_*^T = M\mathbf{p}_*^T$, represents—if it exists—the probability for the particle to be located at a specific node for an infinite amount of time.

In the RWR version, at each iteration, the particle can also restart by jumping to any randomly selected node in the graph, with a defined restart probability, $r \in (0, 1)$. This avoids the walk to be trapped in a dead end, and assures the existence of the stationary distribution (Langville and Meyer, 2004). Moreover, we can restrict the restart of the particle to specific node(s), called seed(s) (Pan *et al.*, 2004). Doing so, the particle will explore the graph focusing on the neighborhood of the seed(s), and the stationary distribution can be considered as a measure of the proximity between the seed(s) and all the other nodes in the graph.

Formally, based on Equation (1), RWR equation can be defined as:

$$\mathbf{p}_{t+1}^T = (1-r)M\mathbf{p}_t^T + r\mathbf{p}_0^T. \tag{2}$$

The vector $\mathbf{p}_0$ is the initial probability distribution. Therefore, in $\mathbf{p}_0$, only the seed(s) have values different from zero. After several iterations, the difference between the vectors $\mathbf{p}_{t+1}$ and $\mathbf{p}_t$ becomes negligible, the stationary probability distribution is reached, and the elements in these vectors represent a proximity measure from every graph node to the seed(s). In this work, iterations are repeated until the difference between $\mathbf{p}_t$ and $\mathbf{p}_{t+1}$ falls below 10^{-10}, as in previous studies (Li and Patra, 2010; Erten *et al.*, 2011; Zhao *et al.*, 2015).

We set the global restart parameter to $r = 0.7$, as in previous studies (Köhler *et al.*, 2008; Li and Li, 2012; Li and Patra, 2010; Smedley *et al.*, 2014; Zhao *et al.*, 2015), for all versions of the RWR algorithm. For the sake of simplicity, we have considered unweighted graphs. However, the extension of the algorithms to weighted graphs is straightforward, and can be achieved by replacing the adjacency matrices by matrices of the weighted edges.

2.2 Random walk with restart on multiplex graphs

2.2.1 Definition

A multiplex graph is a collection of L undirected graphs, considered as layers, sharing the same set of n nodes (De Domenico *et al.*, 2014; Kivelä *et al.*, 2014). Each layer $\alpha = 1, \ldots, L$, is defined by its $n \times n$ adjacency matrix $A^{[\alpha]} = (A^{[\alpha]}(i,j))_{i,j=1,\ldots,n}$, where $A^{[\alpha]}(i,j) = 1$ if node i and node j are connected on layer α, and 0 otherwise (Battiston *et al.*, 2014). We do not consider auto-interactions $(A^{[\alpha]}(i,i) = 0 \,\forall\, i = 1, \ldots, n)$, and v_i^α stands for the node i in layer α. A multiplex graph is characterized by its adjacency matrix:

$$\mathbf{A} = A^{[1]}, \ldots, A^{[L]} \tag{3}$$

and is defined as $G_M = (V_M, E_M)$, where:

$$V_M = \{v_i^\alpha,\ i = 1, \ldots, n,\ \alpha = 1, \ldots, L\},$$

$$E_M = \{(v_i^\alpha, v_j^\alpha),\ i,j = 1, \ldots, n,\ \alpha = 1, \ldots, L,\ A^{[\alpha]}(i,j) \neq 0\} \bigcup$$

$$\{(v_i^\alpha, v_i^\beta),\ i = 1, \ldots, n,\ \alpha \neq \beta\}.$$

2.2.2 RWR-M: extension of RWR to multiplex graphs

The particle can walk from its current node v_i^α to any of its neighbors within a layer, or jump to any node v_i^β with $\beta \neq \alpha$ (De Domenico *et al.*, 2013), and thereby change from one to another layer, as schematically displayed in Figure 1A.

We can thus extend the classical RWR algorithm to a multiplex graph (RWR-M) by building a $nL \times nL$ matrix, A. The matrix A contains the different types of transitions that the simulated particle can follow at each step, and is defined as:

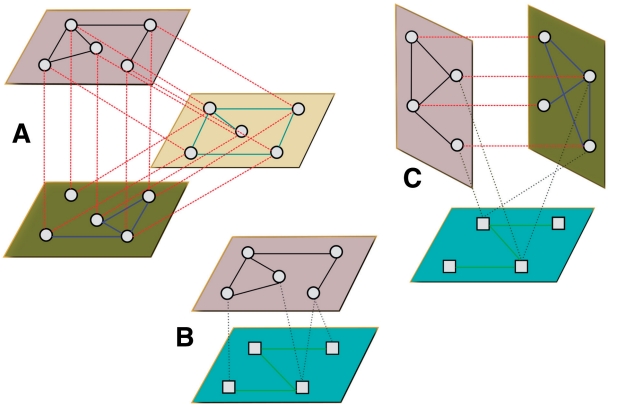

Fig. 1. Multiplex, heterogeneous and multiplex-heterogeneous graphs. (**A**) A multiplex graph composed of three layers. The particle can navigate within each layer or jump to the same node in another layers. (**B**) A heterogeneous graph composed of two graphs. The particle can navigate within each graph or jump to the other graph according to bipartite associations between the two different types of nodes. (**C**) A multiplex-heterogeneous graph

$$A = \begin{pmatrix} (1-\delta)A^{[1]} & \frac{\delta}{(L-1)}\mathbf{I} & \cdots & \frac{\delta}{(L-1)}\mathbf{I} \\ \frac{\delta}{(L-1)}\mathbf{I} & (1-\delta)A^{[2]} & \cdots & \frac{\delta}{(L-1)}\mathbf{I} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\delta}{(L-1)}\mathbf{I} & \frac{\delta}{(L-1)}\mathbf{I} & \cdots & (1-\delta)A^{[L]} \end{pmatrix} \tag{4}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix and $A^{[\alpha]}$ is the adjacency matrix of the layer α, as described in (3). The elements in the diagonal represent the potential intra-layer walks, whereas the off-diagonal elements account for the possible jumps between different layers. The parameter $\delta \in [0, 1]$ quantifies the probability of staying in a layer or jumping between the layers: if $\delta = 0$ the particle will always stay in the same layer after a non-restart step.

Let us denote the transition matrix M obtained by a column normalization of A. Equation (2) in the multiplex case becomes:

$$\bar{\mathbf{p}}_{t+1}^T = (1-r)M\bar{\mathbf{p}}_t^T + r\bar{\mathbf{p}}_{RS}^T \tag{5}$$

where $\bar{\mathbf{p}}_t = [\mathbf{p}_t^1, \ldots, \mathbf{p}_t^L]$ and $\bar{\mathbf{p}}_{t+1} = [\mathbf{p}_{t+1}^1, \ldots, \mathbf{p}_{t+1}^L]$, $t \in \mathbb{N}$, are $n \times L$ vectors representing the probability distribution of the particle in the multiplex graph. These vectors are composed of the probability distributions in every layer. The restart vector, $\bar{\mathbf{p}}_{RS}$, represents the initial probability distribution. We define it as $\bar{\mathbf{p}}_{RS} = \tau \cdot \bar{\mathbf{p}}_0$, where the vector parameter $\tau = [\tau_1, \ldots, \tau_L]$ measures the probability of restarting in the seed(s) of each layer in the multiplex graph. It is to note that it is possible to tune the importance of each layer by modifying the parameter τ.

We established an equal restart probability in all the layers, $\tau = (1/L, 1/L, \ldots, 1/L)$, and we also considered an equal probability for staying in a layer or jumping between the layers, $\delta = 0.5$.

When the stationary probability distribution is reached, every node is associated to L proximity measures, one for each layer of the multiplex graph. We compute the global score for every node as the geometric mean of its L proximity measures. The geometric mean

penalizes nodes with a good score in one layer, but low scores in the remaining layers.

2.3 Random walk with restart on heterogeneous graphs

2.3.1 Definition

A heterogeneous graph contains two graphs with different types of nodes and edges, as well as a bipartite graph containing bipartite associations between them (Lee *et al.*, 2013). Let us consider the graphs $G_V = (V, E_V)$ with $V = \{v_1, \ldots, v_n\}$, $G_U = (U, E_U)$ with $U = \{u_1, \ldots, u_m\}$, and the bipartite graph $G_B = (V \cup U, E_B)$ with $E_B \subseteq V \times U$. The edges of the bipartite graph only connect pairs of nodes from the different sets of nodes, V and U. We can now define a heterogeneous graph, $G_H = (V_H, E_H)$, as:

$$V_H = \{V \cup U\}$$

$$E_H = \{E_V \cup E_U \cup E_B\}.$$

2.3.2 RWR-H: extension of RWR to heterogeneous graphs

Li and Patra (2010) proposed a RWR on a heterogeneous graph. This heterogeneous graph was composed of a PPI network, a disease–disease similarity network, and a bipartite graph containing protein-disease associations. The particle walks on the PPI network, on the disease–disease similarity network, and can also jump between the two networks following the bipartite associations (Fig. 1B). Equations and technical details of the approach proposed by Li and Patra (2010) are described in the Supplementary Methods.

2.4 Random walk with restart on multiplex-heterogeneous graphs

2.4.1 Definition

Let us consider a L-layers multiplex graph, $G_M = (V_M, E_M)$, with $n \times L$ nodes, $V_M = \{v_i^\alpha, i = 1, \ldots, n, \alpha = 1, \ldots, L\}$. Let $G_U = (U, E_U)$ be a graph with m nodes, $U = \{u_1, \ldots, u_m\}$. In order to build a heterogeneous graph composed of G_M and G_U, we need to link the nodes in every layer of the multiplex graph G_M to their associated nodes in the graph G_U, according to their bipartite associations, E_B. Since the same nodes are present in every layer of the multiplex graph, it is necessary to have L identical bipartite graphs, $G_B^{[\alpha]} = (V_M \cup U, E_B^{[\alpha]})$ to define the multiplex-heterogeneous graph. We can then describe a multiplex-heterogeneous graph, $G_{MH} = (V_{MH}, E_{MH})$, as:

$$V_{MH} = \{V_M \cup U\}$$

$$E_{MH} = \{\cup_{\alpha=1,\ldots,L} E_B^{[\alpha]} \cup E_M \cup E_U\}.$$

2.4.2 RWR-MH: extension of RWR to multiplex-heterogeneous graph

We finally extended the RWR algorithm to multiplex-heterogeneous networks (RWR-MH). At a given step, let the particle be at a specific node within a layer of the multiplex graph. At the next non-restart step, the particle can either (i) walk within the same layer or (ii) jump to the same node in a different layer or (iii) jump to the other graph if a bipartite association exists (Fig. 1C).

Let consider a multiplex graph composed of n gene/protein nodes and L-layers, with an adjacency matrix $A_{M(nL\times nL)}$, like the one described in Equation (4). Let also consider a disease–disease similarity graph characterized by its adjacency matrix, $A_{D(m\times m)}$, where m is the total number of diseases. The bipartite graphs with adjacency matrices $B_{(n\times m)}^{1,\ldots,L}$ associate the gene/protein nodes in each layer of the multiplex graph to diseases. These bipartite graphs are identical, we define them as $B_{(n\times m)}$, and construct the bipartite adjacency matrix of the multiplex-heterogeneous graph by sticking $B_{(n\times m)}$ L times.

$$B_{MH} = \begin{pmatrix} B_{(n\times m)} \\ B_{(n\times m)} \\ \vdots \\ B_{(n\times m)} \end{pmatrix}. \quad (6)$$

Then, we can define the global adjacency matrix of the multiplex-heterogeneous graph as $A = \begin{bmatrix} A_M & B_{MH} \\ B_{MH}^T & A_D \end{bmatrix}$, where B_{MH}^T represents the transpose of B_{MH}. From this point, we can proceed in an analogous way to the one describing the RWR on heterogeneous graphs (Supplementary Methods). We define a global transition matrix for the multiplex-heterogeneous network and calculate its components using the same equations. We just have to replace the adjacency matrix of the PPI network, $A_{P(n\times n)}$, by the adjacency matrix of the multiplex network$A_{M(nL\times nL)}$, and the bipartite adjacency matrix, $B_{(n\times m)}$, by the adjacency matrix of the bipartite graph of the multiplex-heterogeneous graph, $B_{MH(nL\times m)}$.

In order to apply the Equation S5 (Supplementary Methods), we have to consider that the vectors $\tilde{\mathbf{p}}_{t+1}$, $\tilde{\mathbf{p}}_t$ and $\tilde{\mathbf{p}}_{RS}$ are now of dimension $((n \times L) + m)$, since the RWR-MH algorithm is ranking n proteins in L different layers and m diseases at the same time. It is to note that it is possible to tune the importance of each network by defining $\tilde{\mathbf{p}}_{RS} = \begin{bmatrix} (1-\eta)\mathbf{u}_0 \\ \eta\mathbf{v}_0 \end{bmatrix}$, where $\mathbf{u}_0$ defines the initial probability distribution of the multiplex graph, as described for the RWR on heterogeneous graphs (Supplementary Methods), and $\mathbf{v}_0$ the initial probability distribution of the disease-disease similarity network.

2.5 Network sources

Network details can be found in Supplementary Methods: sizes and densities (Supplementary Table S1), degree distributions (Supplementary Fig. S1A) and overlaps between nodes and edges (Supplementary Fig. S1B and C). Network figures are represented using Cytoscape (Shannon *et al.*, 2003).

2.5.1 Biological networks

We constructed three biological networks containing genes or proteins as nodes (genes and proteins are here considered equally): a PPI network, a network connecting proteins according to pathway interaction data, and a network in which the links correspond to co-expressed genes (Supplementary Methods). The networks were generated from downloads on November 23 and 24, 2016, and from the source codes available on GitHub. The PPI network contains 12 621 no and 66 971 edges. The Pathway network contains 10 534 nodes and 254 766 edges, and the Co-expression network is composed of 10 534 nodes connected by 1 337 347 edges.

2.5.2 Disease–disease similarity network

Diseases and their associated phenotypes were obtained from the Human Phenotype Ontology Project (HPO) (Köhler *et al.*, 2014), and we constructed a disease-disease similarity network-based on phenotype similarities between every pair of diseases. The similarity value is computed according to the relevance of the shared phenotypes. We estimated the relevance of each phenotype from the

information content (IC) given by its frequency in the HPO database, as proposed by Westbury *et al.* (2015) (Supplementary Methods).

2.5.3 Gene–disease bipartite associations

We connected the nodes in each layer of the multiplex network with the disease–disease similarity network thanks to bipartite gene–diseases associations extracted from OMIM (Hamosh *et al.*, 2005), using biomaRt (Durinck *et al.*, 2009) (downloads December, 2016). We obtain 4496 associations between genes/proteins and diseases.

2.6 Leave-one-out cross-validation

The performances of the different RWR algorithms were evaluated using a LOOCV strategy. Known disease-gene associations from OMIM (Hamosh *et al.*, 2005) and DisGeNET v4.0 (Piñero *et al.*, 2016) were used as a benchmark: for each disease-associated to at least two genes, each associated gene is removed one-by-one, and considered as the *left-out gene*. The remaining genes are used as seed(s) in the RWR algorithms. All the network nodes are then scored and ranked according to their proximity to the seed(s), and the rank of the left-out disease-gene is recorded (Supplementary Methods).

3 Results

Our main goal was to design a RWR algorithm able to exploit multiple biological interaction sources. We first constructed three biological networks: a PPI network, a pathway-derived network and a co-expression network (Materials and methods). These networks can be considered independently as monoplex networks. They can also be merged as an aggregated network, with nodes and edges corresponding to the union of the monoplex networks. The aggregated network is composed of 17 559 nodes and 1 659 084 edges (Supplementary Table S1). Finally, we also studied the three networks as a multiplex network. A multiplex network is a collection of networks considered as layers, sharing the same set of nodes, but in which the edges belong to different interaction categories. In our multiplex network, the layers share the same set of 17 559 nodes, also corresponding to the union of all network nodes. The genes/proteins absent in a layer are added as isolated nodes in this layer.

We also constructed a disease–disease similarity network, in which the nodes correspond to diseases, and the edges to the most significant phenotype similarities between the diseases (Materials and methods). Finally, in order to construct a multiplex-heterogeneous network, we linked the disease–disease similarity network to the multiplex network thanks to bipartite gene–disease associations.

We next devised different RWR algorithms, which each leverage the different networks and combinations thereof, and we compared their efficiencies.

3.1 Random walk with restart on multiplex networks are more efficient than on monoplex networks

The classical RWR algorithm takes as input a monoplex network. Here, we first adapted the RWR algorithm to navigate a multiplex network (RWR-M). Basically, at each step, the particle can walk from one node to another in the same layer, as in a monoplex network, but it can also move to the same node in another layer of the multiplex network (Materials and methods). We compared the performances of the classical RWR and multiplex RWR-M algorithms in retrieving disease-associated genes, thanks to a LOOCV strategy (Materials and methods). For that, we created a test set composed of diseases associated to at least two genes in the set of 4529 protein nodes common to the three networks. This test set contains 273 diseases and 1312 gene–disease associations. For every disease, each of its associated genes is iteratively left-out, and the remaining gene(s) are considered as seed(s) to run the algorithms. We then compared the ability of the different RWR algorithms to retrieve the left-out gene (Fig. 2).

Focusing first on monoplex networks, the worst performance is observed for the classical RWR algorithm applied to the co-expression network. It seems difficult to retrieve known disease-associated genes from a network built from correlations of mRNA expression data alone. The Pathway-derived network achieves the best performance among the monoplex networks, probably because pathways databases are usually built on established biological knowledge and curated. Noticeably, the RWR algorithm is not able to predict disease-associated genes from randomized versions of these biological networks (Supplementary Results).

The RWR-M algorithm, exploiting more than one interaction source in a multiplex framework, performs better than the classical RWR. In particular, despite the low ranking capacities of the co-expression network alone, its integration as a layer in a multiplex framework of two or three layers enhances the performance of the algorithm. Overall, the best results are obtained with the integration of the three network layers (Fig. 2).

3.2 Random walk with restart on multiplex networks are more efficient than on aggregated networks

In a second step, we compared the performances of the RWR on multiplex network (RWR-M) with the classical RWR run on the three networks aggregated as a single monoplex network. In the aggregated network, two proteins can be linked by up to three edges (corresponding to the three network sources), and the particle can choose between these different edges to move from a node to one of its neighbors, as in Li and Li (2012). The ranking ability of RWR-M and classical RWR on the aggregated network are again tested by LOOCV. In this case, we created the test set with diseases associated

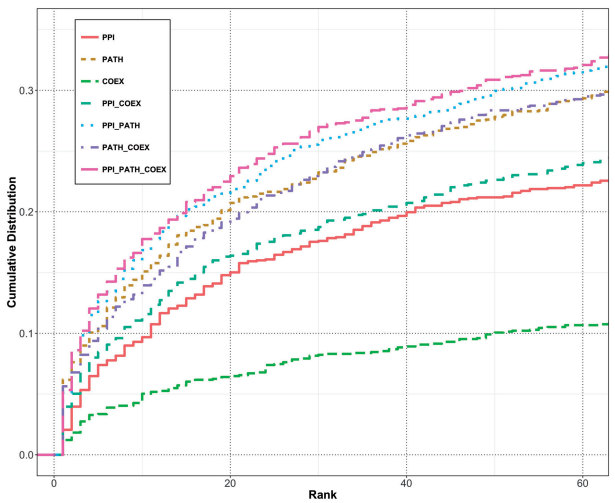

Fig. 2. Cumulative distribution functions representing the ranks of the left-out disease genes in the LOOCV with different RWR algorithms. Classical RWR algorithm is applied to the protein–protein (PPI), Pathway (PATH) and co-expression (COEX) monoplex networks. RWR-M algorithm is applied to combinations of two or three of these networks, considered as layers of a multiplex network

to at least two nodes in the total of 17 559 nodes corresponding to the union of the nodes of the three networks. The test set contains 537 diseases and 2892 gene–disease associations.

The ranks of the left-out disease genes are better with the RWR-M than with the classical RWR on the aggregated network (Fig. 3). The aggregated and multiplex networks use the same biological data and interaction network sources, but the multiplex framework further keeps tracks of the individual topological structures in each network layer.

3.3 Random walk with restart on multiplex-heterogeneous networks are more efficient than on multiplex or heterogeneous networks alone

We previously compared the performances of RWR algorithms on different combinations of networks containing the same nodes but edges belonging to different interaction categories. We now wish to extend these comparisons to heterogeneous networks, i.e. networks containing different sets of nodes, such as genes/proteins and diseases.

We first coded the heterogeneous RWR-H algorithm as proposed by Li and Patra (2010) (Materials and methods). The RWR-H algorithm takes as input a heterogeneous network composed of a PPI network and a disease–disease similarity network. We constructed the disease–disease similarity network by computing the phenotype similarity between a pair of diseases (Materials and methods). The PPI and the disease–disease similarity networks are connected by bipartite gene-disease associations. In the RWR-H algorithm, the particle can move from the PPI network to the disease–disease similarity network thanks to these bipartite associations.

We here compared the ranking capacities of RWR-M and RWR-H by LOOCV. In this case, we created a test set of diseases associated to at least two genes in the set of 12 621 nodes present in the PPI network. The test set contains 242 diseases and 880 gene-disease associations. We can observe first that RWR-M and RWR-H perform better than the classical RWR on the monoplex PPI network (Fig. 4). This stands for other types of heterogeneous networks, built by combining pathway and disease–disease similarity networks, or co-expression and disease–disease similarity networks (Supplementary Fig. S3).

In this context, an algorithm able to execute a RWR on both multiplex-heterogeneous networks is expected to have better performances. Therefore, we extended our RWR-M approach to heterogeneous networks, defining a RWR on multiplex-heterogeneous networks, RWR-MH (Materials and methods). The RWR-MH displays a remarkable amelioration of performances in the prioritization task, since over 45% of the left-out genes are ranked within the top 20 (Fig. 4).

Finally, we further checked the influence of the different parameters involved in the RWR-MH algorithm using the LOOCV strategy. Overall, the RWR-MH is a very robust algorithm since variations in the parameters do not lead to large variations in the ranking performances (Supplementary Results, Supplementary Figs S4 and S5).

3.4 Candidate genes for the undiagnosed Wiedemann–Rautenstrauch syndrome

The Wiedemann–Rautenstrauch neonatal progeroid syndrome (MIM code: 264 090) is characterized by intrauterine growth retardation with subsequent failure to thrive and short stature (Toriello, 1990). Patients also display a progeroid appearance, decreased subcutaneous fat, hypotrichosis and macrocephaly (Kiraz *et al.*, 2012). Only a few published cases have been documented, and to our knowledge, no gene has been described as causative of the syndrome yet.

To illustrate the application of our approach for disease-associated gene prediction, we applied the RWR-MH algorithm using as seed only the WRS disease node. We then considered the top 25 ranked genes as putative candidates for playing a role in WRS (Fig. 5). Many of these top predicted candidate genes, such as *FIG4*, *RNF113A* or *LMNA*, are implicated in diseases directly connected to WRS from phenotype similarities. Mutations in *LMNA* are responsible for the Hutchinson–Gilford progeria syndrome (MIM code: 176 670) and other premature aging syndromes such as

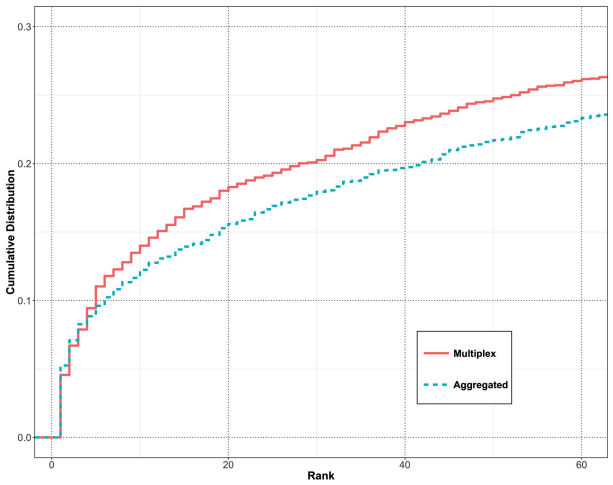

Fig. 3. Cumulative distribution functions representing the ranks of the left-out disease genes in the LOOCV with different RWR algorithms. Classical RWR algorithm is applied on the three networks aggregated as a single monoplex network, and RWR-M algorithm is applied to combinations of the three networks as layers of a multiplex network

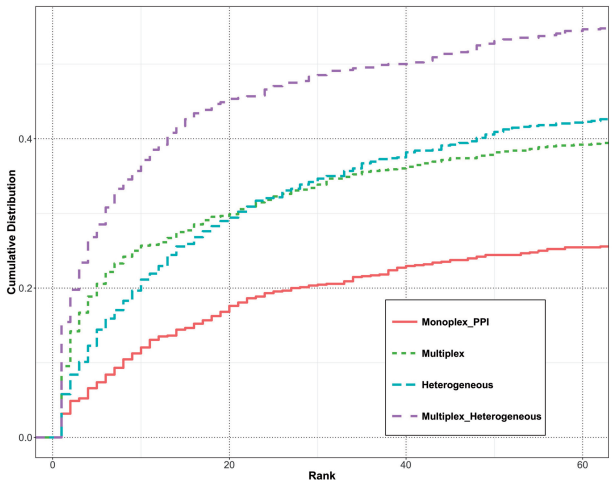

Fig. 4. Cumulative distribution functions representing the ranks of the left-out disease genes in the LOOCV with different RWR algorithms. Classical RWR algorithm is applied to the monoplex PPI network, RWR-M is applied to the combinations of the three monoplex networks as layers of a multiplex network, RWR-H algorithm is applied to the heterogeneous network composed of the PPI network and the disease-disease similarity network, and RWR-MH algorithm is applied the multiplex-heterogeneous network composed of the three-layers multiplex network and the disease-disease similarity network

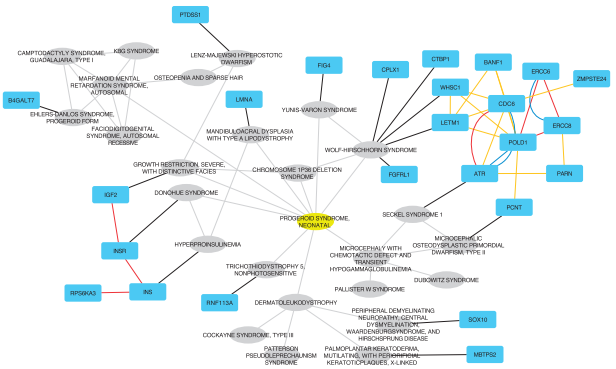

Fig. 5. Network representation of the top 25 ranked genes and diseases when the RWR-MH algorithm is executed using WRS as seed (yellow node). Gray elliptical nodes are diseases; turquoise rectangles are genes/proteins. Black edges are bipartite gene-disease associations from OMIM (Hamosh *et al.*, 2005); grey edges are the similarity links in the disease–disease network; blue edges are PPI interactions; yellow edges are co-expression relationships; red edges are pathway interactions. It is to note that results are represented as an aggregated network only for visualization purposes

Mandibuloacral Dysplasia with type A lipodystrophy (MIM code: 248 370). However, the targeted sequencing of *LMNA* in few WRS patients did not identify mutations (Hou, 2009; Kiraz *et al.*, 2012). The RWR-MH algorithm also top ranked *ZMPSTE24*, which is known to cause severe progeroid syndromes such as Restrictive Dermopathy (MIM code: 275 210) (Navarro *et al.*, 2006). But here also, no mutations were found for this gene in five WRS patients (Hou, 2009).

Another set of interesting candidates is given by the subnetwork composed of the four genes *IGF2*, *INS*, *INSR* and *RPS6KA3*. All these genes participate in the insulin pathway, and are associated to diseases sharing phenotypes with WRS [i.e. Donohue Syndrome (MIM code: 147 670), hyperproinsulinemia (MIM code: 176 730) and severe growth restriction (MIM code: 147 470)]. The insulin pathway is suspected to play a role in WRS (Arboleda *et al.*, 2007). Similarly, a cluster of proteins related to the cell cycle and DNA repair is connected to WRS through the Wolf–Hirschhorn syndrome (MIM code: 194 190), and DNA repair defects are also suspected to be involved in WRS (Hou, 2009).

3.5 Exploring vicinity of *PIK3R1* and SHORT syndrome

SS (MIM code: 269 880) is a rare disease with clinical features defined by its acronym: short stature, hyperextensibility of joints and/or inguinal hernia, ocular depression, Rieger abnormality and teething delay (Gorlin, 1975). However, these phenotypes do not describe the full range of SS phenotypes, and other clinical features include, for instance, partial lipodystrophy and insulin resistance (Avila *et al.*, 2016). Mutations in the *PIK3R1* gene are described as the main cause of SS (Chudasama *et al.*, 2013; Dyment *et al.*, 2013; Thauvin-Robinet *et al.*, 2013).

We applied the RWR-MH algorithm using the *PIK3R1* gene and the SS disease as seeds, and explored the top 25 ranked diseases and genes, along with their interactions and associations (Supplementary Fig. S6). Many of the top ranked diseases recapitulate phenotypes associated to SS. For instance, permanent neonatal diabetes mellitus (MIM code: 606176) accounts for SS phenotypes associated to insulin resistance. Mandibuloacral dysplasia with type B lipodystrophy (MIM code: 608 612) and other diseases associated to lipodystrophy are also top ranked, as well as the growth hormone insensitivity syndrome (MIM code: 262 500) that share with SS the phenotypes related to short stature, among others.

Some of the identified subnetworks are very appealing. For instance, we can observe a loop linking the SS, its associated gene, *PIK3R1*, the Lowe oculocerebrorenal syndrome (MIM code: 309 000) and its associated gene *OCRL*. These two diseases share a noticeable amount of phenotypes, including growth retardation and glucose intolerance. The *PIK3R1* and *OCRL* genes are coding proteins involved in the same pathway: synthesis of phosphatidylinositol phosphates at the plasma membrane (reactome code: R-HSA-1 660 499). Therefore, we can hypothesize a common deregulation of this pathway in the two diseases, leading to shared phenotypes.

Similarly, we can point to the subnetwork containing the *ELN* gene, implicated in the Williams–Beuren syndrome (MIM code: 194 050). Many phenotypes associated to this syndrome are similar to SS and Lowe oculocerebrorenal syndrome. In this case, the *ELN* gene is linked to the *PDGFRB* gene by a co-expression relationship. *PDGFRB* is highly connected to many nodes in the subnetwork, including to *PIK3R1*, by pathway interactions. The co-expression interaction between *PDGFRB* and *ELN* is intriguing because the two genes are, to our knowledge, not described to be involved in the same pathway or process. However, they seem to be regulated by the same microRNA-29 family (Cushing *et al.*, 2015; Zhang *et al.*, 2012). Overall, these results could also allow pointing to other candidate genes predicted to be involved in the SS. This is interesting as, for instance, Dyment *et al.* (2013) did not find any mutation in the *PIK3R1* gene in one of the seven tested patients.

4 Discussion

Physical and functional relationships between genes and proteins are diverse. They are identified or derived from various approaches, each having its own features, strengths and weaknesses. In this context, the integration of different sources of interaction, exploiting data pluralism, is expected to outperform approaches dealing with single networks. Indeed, the combination of different large-scale interaction datasets increases the available biological information, and potentially reduce the bias and incompleteness of isolated sources (Menche *et al.*, 2015).

We and others also hypothesized that the multiplex framework, which retains information on the topology of the individual networks, would perform better as compared to the aggregation of the different interaction sources (Battiston *et al.*, 2014; Didier *et al.*, 2015; Kivelä *et al.*, 2014; Kurant and Thiran, 2006). We have shown previously, for instance, that the multiplex framework is more efficient than network aggregations to extract communities from biological networks (Didier *et al.*, 2015). We extended here the RWR algorithm by designing the RWR-M algorithm able to leverage multiplex networks. The performances of the RWR-M algorithm are clearly improved as compared to previous algorithms navigating monoplex networks, such as RWR on PPI networks (Köhler *et al.*, 2008) or RWR on aggregated networks (Li and Li, 2012). It is particularly interesting to note that even if a monoplex network, such as the co-expression network, displays poor ranking performances isolated, its integration as a layer of a multiplex network leads to an increase of the performance, thereby demonstrating the potential of the RWR-M strategy.

Moreover, we extended our algorithm to deal with multiplex-heterogeneous networks. To this goal, we first built a disease–disease similarity network-based on the IC of the shared phenotypes between every pair of diseases. Previous approaches building

disease–disease networks, such as the ones proposed by (Li and Patra, 2010; Li and Li, 2012), were based on MimMiner (van Driel *et al.*, 2006). MimMiner mines OMIM full-text and clinical synopsis to compute similarity between diseases. Contrarily, our approach is based on the controlled classification of phenotypes in an ontology, and considers both the ontological structure and the frequencies of phenotypes.

Thanks to the LOOCV, we demonstrated that when the new RWR-MH algorithm is applied on this complex multiplex-heterogeneous network, the prioritization results are far better than those of all other versions of the algorithm. We have also demonstrated that the RWR-MH algorithm displays a robust behavior upon variations of the different parameters. This was previously observed for variations in the parameters of a RWR-H algorithm (Li and Patra, 2010; Zhao *et al.*, 2015). The particle keeps exploring the different network layers thanks to the jumps, and still leverage the complementary biological information. This stability is however observed for the average ranking of left-out genes in the LOOCV, but a focused analysis and network representation of the top 25 ranked genes and diseases in real-case applications would reveal variations.

We focused our applications on a multiplex network composed of a PPI, a pathway and a co-expression network. Other biological networks could be collected or constructed from—omics data, and integrated into our multiplex-heterogeneous framework. Functional interactions can be derived, for instance, by connecting genes annotated for the same GOterms (Ashburner *et al.*, 2000). It would also be valuable to include networks with transcription factors— targets genes, non-coding RNAs as well as drug and therapeutic targets.

Funding

A.V. is the recipient of a CIFRE grant 2015/0982 from the French 'Agence Nationale de la Recherche et de la Technologie'. The project leading to this publication has received funding from the Excellence Initiative of Aix-Marseille University - A*Midex, a French 'Investissements d'Avenir' program.

Conflict of Interest: none declared.

References

Arboleda,G. *et al.* (2007) The neonatal progeroid syndrome (Wiedemann-Rautenstrauch): a model for the study of human aging? *Exp. Gerontol.*, **42**, 939–943.

Arroyo,R. *et al.* (2015) Systematic identification of molecular links between core and candidate genes in breast cancer. *J. Mol. Biol.*, **427**, 1436–1450.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.

Avila,M. *et al.* (2016) Clinical reappraisal of SHORT syndrome with PIK3R1 mutations: toward recommendation for molecular testing and management. *Clin. Genet.*, **89**, 501–506.

Battiston,F. *et al.* (2014) Structural measures for multiplex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **89**, 1–16.

Brin,S. and Page,L. (1998) The anatomy of a large scale hypertextual Web search engine. *Comput. Networks ISDN*, **30**, 107–117.

Brohée,S. and van Helden,J. (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, **7**, 488.

Chapple,C.E. *et al.* (2015) Extreme multifunctional proteins identified from a human protein interaction network. *Nat. Commun.*, **6**, 7412.

Chudasama,K.K. *et al.* (2013) SHORT syndrome with partial lipodystrophy due to impaired phosphatidylinositol 3 kinase signaling. *Am. J. Hum. Genet.*, **93**, 150–157.

Cushing,L. *et al.* (2015) Disruption of miR-29 Leads to Aberrant Differentiation of Smooth Muscle Cells Selectively Associated with Distal Lung Vasculature. *PLoS Genet.*,**11**, e1005238–e1005227.

De Domenico,M. *et al.* (2013) Mathematical formulation of multilayer networks. *Phys. Rev. X*, **3**, 1–15.

De Domenico,M. *et al.* (2014) Navigability of interconnected networks under random failures. *Proc. Natl. Acad. Sci. USA*, **111**, 8351–8356.

Didier,G. *et al.* (2015) Identifying communities from multiplex biological networks. *PeerJ.*, **3**, e1525.

Durinck,S. *et al.* (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, **4**, 1184–1134.

Dyment,D.A. *et al.* (2013) Mutations in PIK3R1 cause SHORT syndrome. *Am. J. Hum. Genet.*, **93**, 158–166.

Erten,S. *et al.* (2011) DADA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization. *BioData Min.*, **4**, 19.

Fabregat,A. *et al.* (2016) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **44**, D481–D487.

Franke,L. *et al.* (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.

George,R.A. *et al.* (2006) Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res.*, **34**, e130.

Gorlin,R.J. *et al.* (1975) A selected miscellany. *Birth Defects Orig. Artic. Ser.*, **11**, 39–50.

Hamosh,A. *et al.* (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.

Hou,J.W. (2009) Natural course of neonatal progeroid syndrome. *Pediatr. Neonatol.*, **50**, 102–109.

Kanehisa,M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.

Katsogiannou,M. *et al.* (2014) The functional landscape of Hsp27 reveals new cellular processes such as DNA repair and alternative splicing and proposes novel anticancer targets. *Mol. Cell. Proteomics*, **13**, 3585–3601.

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

Kiraz,A. *et al.* (2012) Wiedemann-Rautenstrauch syndrome: report of a variant case. *Am. J. Med. Genet. A*, **158A**, 1434–1436.

Kivelä,M. *et al.* (2014) Multilayer networks. *J. Complex Netw.*, **2**, 203–271.

Köhler,S. *et al.* (2014) The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.*, **42**, D966–D974.

Kurant,M. and Thiran,P. (2006) Layered complex networks. *Phys. Rev. Lett.*, **96**, 4.

Langville,A. and Meyer,C. (2004) Deeper Inside PageRank. *Internet Math.*, **1**, 335–380.

Lee,S. *et al.* (2013) PathRank: ranking nodes on a heterogeneous graph for flexible hybrid recommender systems. *Expert Syst. Appl.*, **40**, 684–697.

Li,Y. and Li,J. (2012) Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data. *BMC Genomics*, **13** (Suppl. 7), S27.

Li,Y. and Patra,J.C. (2010) Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, **26**, 1219–1224.

Lovász,L. (1993) Random walks on graphs: a survey. In: *Combinatorics, Paul ErdÖs Is Eighty*, Vol. 2. Keszthely, Hungary, pp. 1–46.

Menche,J. *et al.* (2015) Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science*, **347**, 1257601.

Navarro,C.L. *et al.* (2006) Molecular bases of progeroid syndromes. *Hum. Mol. Genet.*, **15**, R151–R161.

Oti,M. *et al.* (2006) Predicting disease genes using protein-protein interactions. *J. Med. Genet.*, **43**, 691–698.

Pan,J-Y. *et al*. (2004) Automatic multimedia cross-modal correlation discovery. In: *KDD '04 Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 653–658.

Piñero,J. *et al*. (2016) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res*., **45**, D833–D839.

Ruepp,A. *et al*. (2010) CORUM: the comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Res*., **38**, D497–D501.

Schwikowski,B. *et al*. (2000) A network of protein-protein interactions in yeast. *Nature Biotechnol*., **18**, 1257–1261.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*., **13**, 2498–2504.

Smedley,D. *et al*. (2014) Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics*, **30**, 3215–3222.

Thauvin-Robinet,C. *et al*. (2013) PIK3R1 mutations cause syndromic insulin resistance with lipoatrophy. *Am. J. Hum. Genet*., **93**, 141–149.

Toriello,H.V. (1990) Syndrome of the month: Wiedemann-Rautenstrauch syndrome. *J. Med. Genet*., **27**, 256–257.

van Driel,M.A. *et al*. (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet*., **14**, 535–542.

Vanunu,O. *et al*. (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol*., **6**, e1000641.

Westbury,S.K. *et al*. (2015) Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. *Genome Med*., **7**, 36.

Xie,M. *et al*. (2015) Network-based phenome-genome association prediction by bi-random walk. *PLoS One*, **10**, e0125138–e0125118.

Zhang,P. *et al*. (2012) Inhibition of MicroRNA-29 enhances elastin levels in cells haploinsufficient for elastin and in bioengineered vessels-brief report. *Arterioscler. Thromb. Vasc. Biol*., **32**, 756–759.

Zhao,Z.Q. *et al*. (2015) Laplacian normalization and random walk on heterogeneous networks for disease-gene prioritization. *Comput. Biol. Chem*., **57**, 21–28.