Gene2DisCo: Gene to Disease Using Disease Commonalities

Marco Frasca

Università Degli Studi di Milano, Dipartimento di Informatica, Via Comelico 39/41, Milano, Italy

Abstract

Objective. Finding the human genes co-causing complex diseases, also known as "disease-genes", is one of the emerging and challenging tasks in biomedicine. This process, termed gene prioritization (GP), is characterized by a scarcity of known disease-genes for most diseases, and by a vast amount of heterogeneous data, usually encoded into networks describing different types of functional relationships between genes. In addition, different diseases may share common profiles (e.g. genetic or therapeutic profiles), and exploiting disease commonalities may significantly enhance the performance of GP methods. This work aims to provide a systematic comparison of several disease similarity measures, and to embed disease similarities and heterogeneous data into a flexible framework for gene prioritization which specifically handles the lack of known disease-genes.
Methods. We present a novel network-based method, Gene2DisCo, based on generalized linear models (GLMs) to effectively prioritize genes by exploiting data regarding disease-genes, gene interaction networks and disease similarities. The scarcity of disease-genes is addressed by applying an efficient negative selection procedure, together with imbalance-aware GLMs. Gene2DisCo is a flexible framework, in the sense it is not dependent upon specific types of data, and/or upon specific disease ontologies.
Results. On a benchmark dataset composed of nine human networks and 708 medical subject headings (MeSH) diseases, Gene2DisCo largely outperformed the best benchmark algorithm, kernelized score functions, in terms of both area under the ROC curve (0.94 against 0.86) and precision at given recall levels (for recall levels from 0.1 to 1 with steps 0.1). Furthermore, we enriched and extended the benchmark data to the whole human genome and provided the top-ranked unannotated candidate genes even for MeSH disease terms without known annotations.

Keywords: gene prioritization, MeSH disease, biological networks, disease semantic similarity

Email addresses: marco.frasca@unimi.it (Marco Frasca), http://frasca.di.unimi.it (Marco Frasca)

1. Introduction

Complex diseases are attributed to multiple heterogeneous causes, and rarely involve abnormalities on a single gene [1]. High-throughput techniques are useful to identify genes likely to be relevant for specific human diseases and can look at genomic regions containing massive amounts of candidate genes. Thus, since the manual verification of individual candidate genes is expensive and time-consuming, computational methods are needed to aid the discovery of disease-genes by ranking candidate genes from the gene sets on the basis of their likelihood of being involved in a certain disease [2]. This problem, called disease-gene prioritization or simply gene prioritization (GP), is nowadays central biomedicine, since knowing the genetic causes of a given disease can help studying effective treatments for the disease itself, in addition to its prevention. This task is challenging and characterized by several issues, including the vast amount and the heterogeneity of the available information, requiring scalable approaches to integrate multiple data sources; the rarity of known disease-genes for most diseases in the existing disease taxonomies (e.g. the OMIM database [3]); the existence of shared profiles among diseases; the possible association of the same gene with multiple genetic abnormal phenotypes. These difficulties have led to the introduction of numerous categories of study to solve the GP problem.

Text mining approaches attempted in discovering common patterns between genes/proteins and diseases by scanning the literature to find co-occurrences statistically significant [4, 5, 6]. Other proposed tools to discover candidate genes for human genetic diseases relied on *genome wide association studies* (GWAS), asserting that multiple, common small-risk variants interact to cause common diseases [7, 8, 9, 10]. Each study can look at hundreds or thousands of loci at the same time; nevertheless, this approach tends to produce many false-positive results (that is detected candidate genes which are not really involved in the disease etiology), and the experimental validation of candidate genes, for instance through resequencing, pathway or expression analysis, is still expensive and time requiring [11].

Other works leveraged whole exome sequencing to capture all exonic and flanking sequences and to include probes targeting microRNA and other sequences of interest [12]. These studies have reported a successful molecular diagnosis in up to 25% of cases in large cohorts of unselected, consecutive patients [13]. Phenotype-driven analyses of exome data have also been investigated with the aim of filtering out common variants and those deemed to be non-pathogenic [14].

Within the Network medicine context, alternative approaches have widely employed the so called *guilt-by-association* (GBA) principle, in which candidate disease genes are ranked by exploiting the assumption that similar genes tend to share similar diseases [15]. They base on gene networks, in which nodes are genes and connections represent precomputed functional relationships among genes, like protein-protein interactions [16], or transcriptional co-expression

regulation [17]. Network-based methods differ from each other in the way they exploit disease-genes and their direct connections, ranging from protein-protein interaction network analysis and semi-supervised graph partitioning [17, 18], to flow propagation [19], random walks [20], kernelized score functions [21], Gaussian fields and Harmonic functions [22], multiple kernel learning [23], regression trees on mutual information gene networks [24] and network weights adjustment according to a given disease [25]. These methods tend to prefer better characterized genes and/or diseases, due to their need of already discovered disease-genes. Thus, in particular where little prior knowledge about diseases is available, the accurate prioritization of putative disease-genes remains a challenge. Actually, thousands of known genetic diseases in the OMIM (Online Mendelian Inheritance in Man) database have no established gene-disease associations [3].

To partially address this issue, fewer works used tissue-specific expression patterns on the hypothesis that genes responsible for a tissue(s)-specific phenotype are expected to be more expressed in affected than unaffected tissues [26, 27, 28]. However, not considering and integrating potentially complementary evidence coming from other heterogeneous data sources may be a limitation [29]. For instance, interacting pairs of candidate proteins and proteins encoded by known phenotype susceptibility genes (a type of data which also targets rare alleles) can be crucial in prioritizing genes, since two proteins involved in the same biological (dys)function often interact [30].

For these reasons, several research groups have adopted integrated methodologies to exploit at the same time multiple heterogeneous sources, ranging from functional profiles and expression quantitative trait loci to protein complexes and genetics or physical interactions [26, 29, 31]. General approaches to integrate different information sources rank candidate genes with reference to any single source using various metrics, to combine the obtained ranks in an overall rank [15]. For network-based information sources, an integrated network is constructed by combining the topology of each network into a consensus network more informative and with larger coverage [32].

Finally, a promising category of GP methods has focused on the analysis of common characteristics among different diseases [33, 34], discovering sets of interacting proteins and molecular pathways often shared by multiple diseases [35]. The main benefit of such an approach is that 'transferring' information from comparable diseases provides researcher with additional predictive information to prioritize genes, and in particular for less studied abnormalities [36], thus leading to new therapeutic treatments not previously considered. For instance, shared molecular connections between diabetes and dementia are now fueling research into the possible use of insulin to treat Alzheimer's disease [37].

Most GP methods mentioned above tend to focus on some of the issues characterizing the gene prioritization problem, thus limiting their generalization abilities to specific data and/or settings. For instance, the class imbalance problem characterizing GP is neglected by most existing approaches: the availability of a very low number of annotated genes (the positive instances) creates a disproportion between positive and negative instances (non-causative genes)

which leads to a decay of performance for imbalance-unaware classification algorithms [38, 39]. Indeed, employing imbalance-aware algorithms obtained successful results in similar contexts, like the gene function prediction [40, 41]. Moreover, the majority of proposed techniques tend to prioritize disease-genes focusing on single diseases, limiting in this way the amount of prior information adopted as input and the effectiveness of the provided predictions [42].

In this study, we present a novel network-based method, named Gene2DisCo (**G**ene to **D**isease using **Dis**ease **Co**mmonalities), for prioritizing candidate disease-genes by taking into account all the issues of the GP problem. A unique feature of Gene2DisCo is indeed the synergy of data source integration, imbalance-aware learning, and exploitation of disease similarities. In particular, our approach extends a couple of methods recently proposed to prioritize genes [43, 44], by embedding an effective strategy to transfer learning from hundreds of diseases sharing common features. We argue that exploiting shared features among diseases is fundamental for achieving high accuracy in prioritizing disease-genes. To this end, we compared several state-of-the-art semantic similarity measures among nodes in a hierarchy (like MeSH terms), including measures depending on the hierarchy structure, measures depending on the disease information (known disease-genes), and their combination. In addition, by leveraging diseases similarities we have been able in inferring putative associations for more than 800 MeSH diseases, including around 400 diseases with no discovered genes, thus overcoming the main limitation of existing methods.

We validated our method on benchmark data including nine human gene networks, containing 8449 genes and 708 MeSH disease terms, with statistically significant improvements over benchmark methods. In order to supply reliable novel gene-disease associations, we then constructed an enriched data set by integrating existing networks with the STRING network, version 10.0 [45], and by downloading recent human gene association for MeSH diseases from the Comparative Toxicogenomics Database (CTD) [46]. The obtained data contain 19112 human genes and cover diseases at different levels of specificity and spanning different general categories of MeSH diseases.

Overall, the main contributions of this study are summarized as follows:

(i) A novel and flexible framework Gene2DisCo to prioritize genes not dependent on specific sources of data and/or of gene-disease annotations and composed by the following modules (see Figure 1): a module to integrate different and heterogeneous network-based data sources; 2) a module to embed information about diseases similar to the disease under study; 3) a module to infer imbalance-aware predictions on the basis of input data, including data provided by modules 1 and 2.

(ii) A comparative analysis of different methods to compute pairwise semantic similarities among diseases and a quantitative evaluation of their impact on gene prioritization.

(iii) One of the widest data set for prioritizing genes, involving gene-MeSH disease associations covering hundreds of diseases, and extended to the whole human genome.

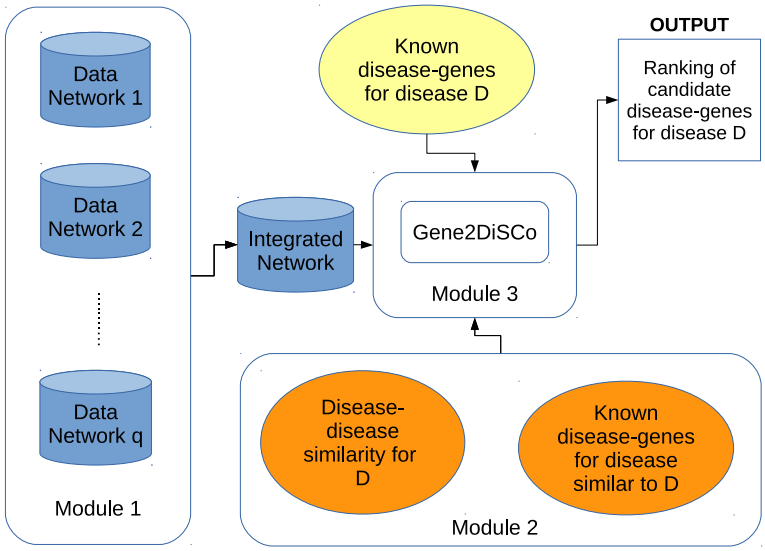

Figure 1: Gene2DisCo functional modules: integration of different sources about genes; retrieval of disease-disease commonalities and of diseases-genes for diseases sharing commonalities with D; inference of the gene list so as to rank higher genes more likely to be involved in the etiology of D.

(iv) The inferred predictions for around 800 MeSH diseases, including 348 diseases without known causal genes.

This paper is organized as follows. In Section 2 we describe the benchmark data used to validate Gene2DisCo and the refined data set embedding the STRING network. Section 3 discusses the GP problem, the data integration procedure, the disease similarity measures adopted, and the generalized linear model proposed to solve the GP problem. The experimental setting and validation, and the top-ranked unannotated genes for the considered MeSH diseases are presented in Section 4. The conclusion remarks end the paper.

2. Materials

We present here the networked data sources adopted to prioritize genes. In particular, we first describe the benchmark data set utilized to validate our method, and then we propose an enriched data setting extended to the whole human genome and including more recent gene-disease associations.

2.1. Benchmark data

We follow the benchmark setting proposed in [47], which includes nine human gene networks covering 8449 genes (or a subset of them), and the associations of such genes with 708 selected MeSH disease terms, downloaded from the CTD database. MeSH is a vocabulary thesaurus, being controlled by the National Library of Medicine to index MEDLINE documents, which consists of a set of description terms organized in a hierarchical structure (called MeSH trees), where more general terms appear at nodes closer to the root and more specific terms appear at nodes closer to leaves [48]. Each MeSH node is represented by a tree number, which indicates the position of the node in MeSH tree. Furthermore, every MeSH heading may correspond to more than one node in different MeSH trees (or subtrees). Finally, each tree in the MeSH hierarchy belongs to a semantic category, among which the headings considered in this work fall into categories *C - Diseases* and *F - Psychiatry and Psychology*.

The benchmark nets cover different types of information: functional interactions, transcriptional co-expression/regulation and localization, gene expression profiles, genes-chemicals relationships, protein-protein physical and genetic interactions (2 networks), Gene Ontology (GO) [49] semantic similarity (3 networks). The benchmark MeSH terms posses between 5 and 200 annotated genes, thus avoiding to consider both diseases with too few available information and diseases with a large and heterogeneous set of associated genes (since a gene annotated with a term is also annotated with its ancestors). A brief description of every gene network is supplied below.

Functional interaction network – finet. Composed of 8441 selected proteins, this net contains protein-protein functional binary interaction inferred through a Naive Bayes classifier, trained by using information coming from expert-curated biological pathways and from other non curated sources, such as gene co-expression and protein domain interaction [50].

Human net – hnnet. Starting from 21 large-scale genomics and proteomics data sets from four species, in [51] a functional gene network is integrated by including distinct lines of evidence, spanning human mRNA co-expression, protein-protein interactions, protein complex, and comparative genomics data sets, in combination with similar lines of evidence from orthologs in yeast, fly, and worm.

Cancer module network – cmnet. A gene-gene network of 8849 genes, in which two genes are connected if they share at least one of the 263 biological and clinical conditions considered in [52], where authors collected expression profiles in different tumors and the related behavior of gene modules.

Gene chemical network – gcnet. A network of 7649 genes based on gene-chemical interactions available at the CTD database.

BioGRID database network – dbnet. A net based upon direct physical and genetic interactions obtained from BioGRID (v. 3.2.96 January 2013) [53], and including 8449 proteins.

BioGRID projected network – bgnet. This net relies on the construction of a bipartite graph by using physical and genetic BioGRID interactions between the benchmark 8449 genes and all the human genes available in BioGRID. The net is pruned to the benchmark genes by inserting and edge between two genes if they share at least on neighbour in the bipartite graph.

Semantic similarity–based networks: bpnet, mfnet and ccnet. For every branch (*biological process*, *molecular function* and *cellular component*) of GO taxonomy, a net is constructed by considering the GO terms the 8449 genes are annotated with, and by setting an edge between two genes if they share at least one annotation in the corresponding GO branch. The edge weight is the maximum Rensik semantic similarity [54] between all the terms for which the two genes are both annotated.

2.2. Refining benchmark data

We then constructed a novel data set, including a larger set of human genes and current and more reliable gene-disease associations. We retrieved the input network from STRING database, version 10.0 [45], including all the available genes and their pairwise interactions to maximize the coverage. This network is highly informative, since STRING curators already merge several sources of data, including protein homology relationships from different species. Moreover, in order to enrich the information encoded into the obtained network, we also integrated by unweighted sum the networks described in Section 2.1 (on the union of genes, see Section 3.2), excluding the network *gcnet* which is biased toward MeSH disease terms, as explained in [44]. The final network covers 19112 genes and has around 5.9 millions of distinct edges.

We downloaded the gene-disease associations from the CTD database (04.17), obtaining is a set of 470 MeSH disease terms with 2-200 positive genes, of which 437 posses at least 5 known associations. We excluded diseases with more than 200 associations in order not to work with too generic terms in the MeSH ontology; on the other side, we included diseases with less than 5 associations to work even with more specific (and/or less studied) diseases, exploiting the ability of Gene2DisCo to transfer information from similar diseases. The enriched STRING net and the corresponding MeSH associations are available at `http://frasca.di.unimi.it/data/gene2disco/`.

3. Methods

This section is devoted to illustrate the automated methodologies we adopted to solve the gene prioritization problem. First, we introduce some preliminary definitions, and provide the details of a state-of-the-art method, *NWGP*, recently proposed for prioritizing genes; then, we describe our proposed methodology extending *NWGP* to embed disease pairwise similarities into the model.

3.1. The problem

The gene prioritization problem can be cast into the framework of learning node labels in partial labeled graphs. In this context, a gene network can be represented through an undirected weighted graph $G = (V, \boldsymbol{W})$, where $V = \{1, 2, \ldots, n\}$ is the set of vertices corresponding to genes, and $\boldsymbol{W}$ is the $n \times n$ weight matrix, where each element $W_{ij} \in [0, 1]$ represents some notion of functional similarity between vertices i and j. Vertices in V can be partitioned into two subsets: $L \subset V$ containing instances labeled according to a specific class (MeSH disease term in our context), and its complement $U = V \setminus L$ which, including unlabeled instances, represents the object of our inference. The set of labeled vertices is further partitioned, according to the labeling vector $\boldsymbol{l} \subset \{-1, 0, 1\}^{|V|}$, in the sets $L_+ := \{i \in L | l_i = 1\}$ and $L_- := \{i \in L | l_i = -1\}$ of the positive and negative vertices respectively. Here we assume unlabeled instances posses label 0, denoting the notion of 'no available information'. Furthermore, the labeling $\boldsymbol{l}$ is subjected to a severe imbalance in favor of negative instances, that is $|L_+| \ll |L_-|$.

The gene prioritization problem, which falls in the realm of *Label Prediction in partially labeled Graphs* (LPG), consists in learning a function $\phi : U \to \mathbb{R}$ ranking unlabeled genes/vertices so as to assign higher positions to instances candidate for the positive class. The function ϕ, suitably thresholded, can be then used to assign positive or negative labels to unlabeled vertices.

3.2. Graph integration

The $q = 9$ graphs $G^{(1)} = (V^{(1)}, \boldsymbol{W}^{(1)}), \ldots, G^{(q)} = (V^{(q)}, \boldsymbol{W}^{(q)})$ corresponding to the nets described in Section 2 have been integrated in a unique composite graph $G = (V, \boldsymbol{W})$ by taking the union of all genes, as done in the benchmark setup. After having extended each graph to the union V of vertices by adding zeros in the missing entries of the corresponding adjacency matrix, to integrate them we adopted the *unweighted sum* integration, which performed better than other unweighted schemes used in the benchmark setting. Informally, it consists in averaging the adjacency matrices of all available graphs, that is

$$\boldsymbol{W}^* = \sum_{k=1}^{q} \boldsymbol{W}^{(k)} / q \ .$$

The Laplacian normalization is finally applied to the integrated network $\boldsymbol{W}^*$ obtaining the matrix $\boldsymbol{W} = \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{W}^* \boldsymbol{D}^{-\frac{1}{2}}$, where $\boldsymbol{D}$ is the diagonal matrix of node weighted degrees.

3.3. Disease semantic similarities

As mentioned in Section 2.1, MeSH disease terms are structured as a hierarchy and thereby not independent from each other, and accordingly a disease may share more features with some diseases and less or none with the remaining diseases. For this reason, here we investigate several ways to compute pairwise

similarity measures among diseases. In the following we recall the main state-of-the-art approaches for computing similarities among nodes in a hierarchy, dividing them in two categories: *path-based*, using solely the hierarchy structure, and *information-based*, using the information at each disease (the genes known to be involved in the disease) alone or along with the disease hierarchy information.

Basic definitions. In general, a disease hierarchy may posses different structures (forest of trees like MeSH hierarchy, direct acyclic graph like the Human Phenotype Ontology [55], etc.) and can be represented by a graph $H = (C, E)$, where $C = \{1, 2, \ldots, m\}$ is the set of nodes/headings, E is the set of hierarchical relationships among nodes. The aim is determining a matrix $\mathbf{\Psi} \in \mathbb{R}^{m \times m}$ of pairwise node similarities. Before describing several ways proposed in the literature to compute $\mathbf{\Psi}$, we introduce some definitions used throughout the paper. By $\text{anc}(k) \subset C$ we denote the set of ancestors of node $k \in C$, and by lev_k we denote the level of k in the hierarchy, intended as the number of nodes on the maximum length path from a root node. The level of a root node is thereby 1. To facilitate the following discussion, we assume a unique root in present in H. As in a graph we might have multiple roots, we assume in this case a dummy node is added as parent of root nodes. Moreover, we use $\nu(k)$ to denote the frequency of positive instances for node k and $\boldsymbol{l}^{(k)}$ to denote the labeling vector for node/disease k (see Section 3.1). Since parent-child connections describe specializations of the parent node, a positive instance for a node k must be positive even for any $r \in \text{anc}(k)$. Hence it holds that $\nu(k) \leq \nu(r)$ for any $r \in \text{anc}(k)$. Finally, we denote by $\text{MA}(k, r)$ the common ancestor of nodes $k,\ r \in C$ whose frequency $\nu(\text{MA}(k, r))$ is the lowest among all ancestors of both k and r.

We describe now the main approaches proposed in the literature for computing semantic similarities in a hierarchy, diving them in two main categories: *path-based similarity* measures, exploiting solely the hierarchy structure, and *information-based similarity* measures, leveraging the information content of nodes.

Path-based similarity measures. A first category of methods compute similarity between two nodes as a function of the length of the path linking the nodes, and/or on their position in the hierarchy. In particular we distinguish the following:

SP: Shortest Path [56]. This measure is based on the simple observation that the closer two nodes $k, r \in C$, the more similar they result. This measure is computed through the ratio of the shortest path $sp(k, r)$ between k and r, and the maximum path length between two nodes in C:

$$\psi_{SP}(k, r) = 1 - \frac{sp(k, r)}{\max_{s,q \in C} sp(s, q)} \ . \tag{1}$$

Thus, $0 \leq \psi_{SP}(k,r) \leq 1$, and the most distant nodes have similarity 0. Here $sp(k,r)$ is the function associating nodes k and r with the length of their shortest path (the path weight is the number of edges on the path).

WL: Weighted Links [57]. This measure extends the *SP* measure by assigning weights to paths depending on the *depth* of nodes forming the path. The weight of a node $k \in C$ is the reciprocal of its level in the hierarchy. Thus:

$$\psi_{WL}(k,r) = 1 - \frac{wl(k,r)}{\max_{s,q \in C} wl(s,q)} \ . \tag{2}$$

where $wl(k,r)$ is the function associating nodes k and r with the weight of their shortest path $(s_1, s_2, \ldots, s_{n_{kr}})$, that is $wl(k,r) = \sum_{i=1}^{n_{kr}} 1/\mathrm{lev}_{s_i}$.

WP: Wu and Palmer [58]. Considering the nearest common ancestor $\mathrm{NA}(k,r)$ of two nodes $k, r \in C$, it computes:

$$\psi_{WP}(k,r) = \frac{2\,\mathrm{lev}_{NA(k,r)}}{\mathrm{lev}_k + \mathrm{lev}_r} \ . \tag{3}$$

The lowest similarity between two nodes is obtained when their common ancestor is the root node, and they posses a high level.

LC : Leacock and Chodorow [59]. This measure differs from ψ_{SP} in sense that it takes the logarithm of the resulting score plus a pseudocount:

$$\psi_{LC}(k,r) = 1 - \frac{\log(sp(k,r)+1)}{\log(\max_{s,q \in C} sp(s,q)+1)} \ . \tag{4}$$

Li : Li et al. [60]. Intuitively and empirically derived, it combines the shortest path and the level of the nearest common ancestor in a non-linear function, as follows:

$$\psi_{Li}(k,r) = e^{-\alpha \psi_{SP(k,r)} \frac{e^{\beta\,\mathrm{lev}_{\mathrm{NA}(k,r)}} - e^{-\beta\,\mathrm{lev}_{\mathrm{NA}(k,r)}}}{e^{\beta\,\mathrm{lev}_{\mathrm{NA}(k,r)}} + e^{-\beta\,\mathrm{lev}_{\mathrm{NA}(k,r)}}}} \ , \tag{5}$$

where α and β are real parameters weighting the contribution of the shortest path length and level respectively. According to [60], we set $\alpha = 0.2$ and $\beta = 0.6$.

Information-based similarity measures. Let $-\log(\nu(k))$ be the information content of node $k \in C$. Several measures of similarity have been proposed in the literature exploiting information contents of nodes, with or without considering the hierarchy structure. Here we describe those adopted in this work.

Lord : Lord et al. [61]. This measure is based on the frequency of the common ancestor with the lowest frequency:

$$\psi_{Lord}(k,r) = 1 - \nu(\mathrm{MA}(k,r)) \ . \tag{6}$$

Hence two nodes tend to be more similar when their minimum common ancestor has a low frequence (thus more specific).

Resnik : Resnik [54]. The similarity of two nodes is the information content of their minimum common ancestor:

$$\psi_{Resnik}(k,r) = -\log \nu(\mathrm{MA}(k,r)) \ . \tag{7}$$

Lin : Lin [62]. Given two nodes $k, r \in C$, their similarity is the ratio of the information content of their closest common ancestor $\mathrm{MA}(k,r)$ (their commonalities) and the sum of the information content of k and r (all information needed to describe k and r):

$$\psi_{Lin}(k,r) = \frac{2\log \nu(\mathrm{MA}(k,r))}{\log \nu(k) + \log \nu(r)} \ .$$

JC : Jiang and Conrath [63]. The similarity between nodes $k, r \in C$ is based on the following distance:

$$\mathrm{Dist}(k,r) = -\log \nu(k) - \log \nu(r) + 2\log \nu(\mathrm{MA}(k,r))$$

The distance Dist is usually used to get the corresponding similarity as follows:

$$\psi_{JC}(k,r) = \frac{1}{Dist(k,r)+1} \ .$$

Jaccard: Jaccard [64]. Given nodes $k, r \in C$, their Jaccard similarity is:

$$\psi_{Jaccard}(k,r) = \begin{cases} \dfrac{\left|\{i \in L | l_i^{(k)} = 1 \wedge l_i^{(r)} = 1\}\right|}{\left|\{i \in L | l_i^{(k)} = 1 \vee l_i^{(r)} = 1\}\right|} & \text{if } \{i \in L | l_i^{(k)} = 1 \vee l_i^{(r)} = 1\} \neq \emptyset \\ 0 & \text{otherwise.} \end{cases}$$

This is the ratio between the number of genes that are positive for both nodes and the number of genes that are positive for at least one node. The higher the number of shared genes, the higher the similarity (up to 1). When two nodes do not share any gene, their similarity is zero. In a disease hierarchy, diseases with many positive instances are usually closer to the root (less specific). In this case, the denominator of $\psi_{Jaccard}$ tends to reduce the similarity between the two diseases as opposed to the case in which diseases have a small number of associated genes. Indeed, sharing positives between two specific diseases (closer to leaves) is more informative than sharing positives between two more general diseases (closer to the root).

3.4. Algorithmic scheme

Recently, an effective approach *NWGP* to gene prioritization exploiting generalized linear models (GLMs) for solving the *LPG* problem has been proposed [44]. In this method, which extends a previous approach *WGP* [43],

the response variable decrees the membership to either the positive or negative class (causative and non causative genes). We retain GLMs suitable for *LPG* mainly for three reasons: (i) the need to keep the computational burden low, so as to efficiently deal with the large size of processed data; (ii) their ability to cope with unbalanced labelings (like in our context) by assigning different misclassification costs to positive and negative instances when learning the model; (iii) the easy interpretability of the results characterizing GLMs.

NWGP algorithm adopts a selection of appropriate features associated with vertices/genes able to speed-up the process and to retain the information for an accurate classification: it adopts 2 features, selected in order to suitably represent the local imbalance information at each vertex. Nevertheless, while this algorithm predicts gene-disease associations for a disease independently from the other diseases, here we argue that the semantic relatedness among diseases may strongly help the process of gene prioritization.

In the following we first shortly describe the *NWGP* approach, then we focus on our novel approach to overcome some limitations of *NWGP* by embedding in the model the disease-disease relatedness information.

3.4.1. NWGP:Negative Selection for Weighted Gene Prioritization

The first step of this algorithm associates with each vertex $i \in L$ a point $\Delta_i = (\Delta_i^+, \Delta_i^-)$ in the plane, where:

$$\Delta_i^+ = \sum_{j \in L_+} w_{ij}, \qquad \Delta_i^- = \sum_{j \in L_-} w_{ij}. \tag{8}$$

This projection allows to avoid the *curse of dimensionality* problem, since the projected space only has two dimensions, and to handle the class imbalance problem by assigning different misclassification costs to positive and negative points during the learning phase. Then a negative selection procedure is applied to retain solely the reliable negative points during the learning (we recall that in the MeSH taxonomy solely positive associations are stored). This procedure performs a fuzzy clustering of positive points $P_+ = \{\Delta_i | i \in L_+\}$ [65], and then it scores negative points $P_- = \{\Delta_i | i \in L_-\}$ according to their maximum membership to the h found clusters:

$$\sigma(\Delta) = \max_{1 \leq k \leq h} \left\{ \frac{1}{d(\Delta, c_k)^{\frac{2}{\alpha - 1}}} \right\} \left(\sum_{j=1}^{h} \frac{1}{d(\Delta, c_j)^{\frac{2}{\alpha - 1}}} \right)^{-1}, \tag{9}$$

where $c_1, \ldots, c_h$ are the cluster centroids, d denotes the distance function used by the fuzzy C means (FCM) algorithm [66], and α is a free fuzzification parameter. In the rest of this paper we adopt the $L3$ distance, which performed slightly better than the other distances tested in [44].

The scores associated through σ to negative points are used to: 1) discard negatives whose membership is higher than a given threshold τ (set as the z-quantile of the empirical distribution of the memberships), obtaining novel

sets of negative instances $\bar{L}_- = \{i \in L_- | \sigma(\Delta_i) > \tau\}$ and of negative points $\bar{P}_- = \{\Delta_i | i \in \bar{L}_-\}$; 2) assign weights to the training instances as follows:

$$\omega_i = \begin{cases} \Delta_i^+ / \sum_{j \in L_+} \Delta_j^+ & \text{if } i \in L_+ \\ (1 - \sigma(\Delta_i)) / \sum_{j \in \bar{L}_-} (1 - \sigma(\Delta_j)) & \text{if } i \in \bar{L}_- \end{cases} \tag{10}$$

Authors learned several GLMs by using the computed weights (10) to separate the selected positive and negative points $P_+ \cup \bar{P}_-$, with the *complementary log-log* link function showing the best performances (see [44] for details). Following the suggestions in [44], in this work we set $z = 0.5$, that is half of negative points is discarded.

3.4.2. Learning limitations of NWGP

Intuitively, through the projection shown in Eq. (8), the more a vertex i is functionally similar to positive labeled vertices, the higher the value of Δ_i^+, and analogously for the contribution given by negative labeled vertices to the second coordinate. Remembering the one-to-one correspondence between genes and vertices, with this projection the aim is to find a bipartition of vertices in L which concentrates positive points mostly toward the rightmost lower region of the first quadrant, and the negative points in the rest of the same area. The meaningfulness of such partition relies in turn on the extent to which the set of positive (resp. negative) genes satisfies the two clustering properties of showing high functional similarity with the other genes in the same class and low similarity with genes negatively (resp. positively) labeled as well.

Unfortunately, as evident from Figure 2(a-b), depicting two typical instances of the *LPG* problem downline of the above projection (MeSH ID *D003555*, *C565836* respectively), there is no clear distinction between positive and negative points. Actually, the former are indeed located toward the right extreme of the horizontal axis, contrarily to the negative instances which are distributed in a wider region of the plane; however, both are so intimately mixed up that, in principle, it would be very difficult to distinguish the two classes with a generalized linear model, unless non negligible effects of the two-way interaction between Δ^+ and Δ^- are observed. Two this end, in Figure 2(c-d) we enriched the same instances shown in Figure 2(a-b) by adding a third coordinate in terms of the interaction term $\Delta^+\Delta^-$. Nevertheless, we do not observe a significant improvement of the ability in the discriminating positive and negative points in the new space, since the shape became glancing convex but points still mixed up. This means that the newly added coordinate would not significantly help the classification ability of the model.

We further show this in the following discussion. With reference to labeled vertex i, let us denote with x_{i1} and x_{i2} the coordinates Δ_i^+ and Δ_i^-, and with $y_i := y(x_{i1}, x_{i2})$ its membership to the positive ($i \in L_+$) or negative class ($i \in L_-$), for $i \in V$. We adopt in this example a linear regression algorithm for the straightforward interpretability of its coefficients. Denoting with $\boldsymbol{\alpha} =$

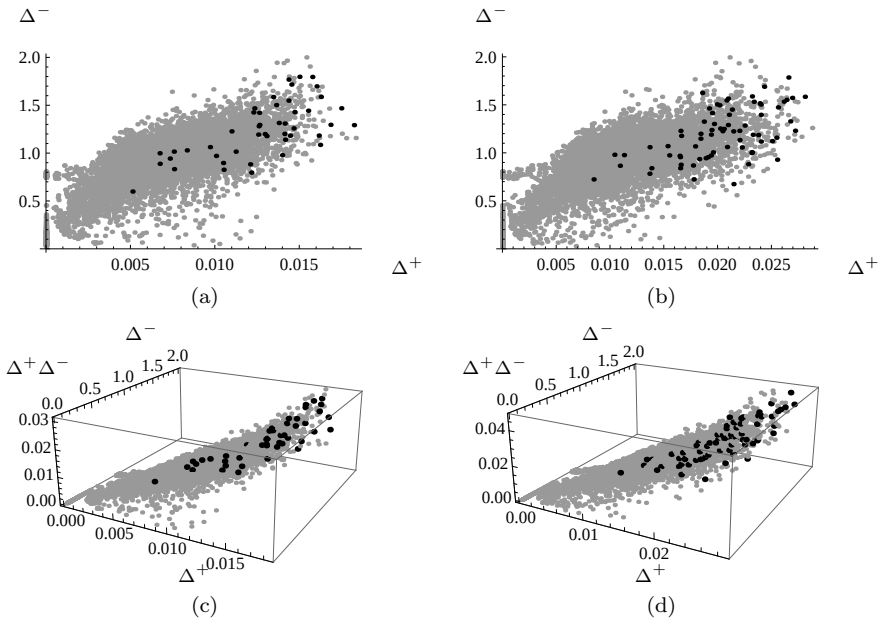

Figure 2: (a-b) Two typical projections in the Δ coordinate space obtained by applying Eq. (8) to two instances of *LPG* problem (namely two MeSH diseases in the benchmark data set described in Section 2.1). Gray and black points correspond to negative and positive labeled vertices respectively. (c-d) The same instances represented in a three-dimensional space, where each node i has coordinates $(\Delta_i^+, \Delta_i^-, \Delta_i^+\Delta_i^-)$.

$(\alpha_0, \alpha^+, \alpha^-, \alpha^\pm)$ the vector of regression coefficients, Figures 3(a-b) show how the linear model:

$$y(x_1, x_2) = \alpha_0 + \alpha^+ x_1 + \alpha^- x_2 + \alpha^\pm x_1 x_2 \tag{11}$$

is able to discriminate between positive and negative labeled nodes in the two lead examples considered so far, distinguishing their labels through a different color. To visually understand the discriminative power of the model, in Figures 3(c-d) we show the curves representing the best learned separators, where points located above (resp. below) the separator will receive a positive (resp. negative) label; it is evident the suboptimal nature of the separating profile, with the majority of positive points wrongly classified. We recall that in this context positives carries almost all the available information.

In conclusion, even adding a the further feature $\Delta^+\Delta^-$ $(x_1 x_2)$, projected points still were located so as to make difficult the separation of positive and negative classes. This is likely due to the fact that the third feature we added is related to the other two features; in the next section we show how the semantic contribution of ontologically related diseases can be used to define a novel

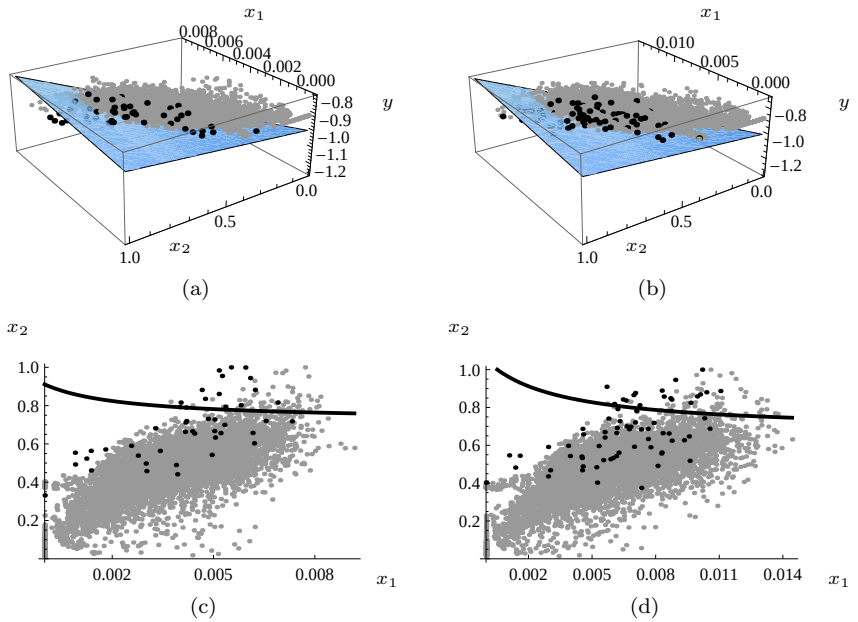

Figure 3: Linear fit of the points shown in Figure 2 (c-d) in the three-dimensional space having coordinates (x_1, x_2, y) (a-b), and in the plane (x_1, x_2) (c-d).

third feature for the model, able in increasing the discriminative power while preserving the computational efficiency.

3.4.3. Generalized linear models embedding disease relatedness: Gene2DisCo

We introduce in this section *Gene2DisCo* (**Gene** to **Dis**ease using Disease **Co**mmonalities), an algorithm extending *NWGP* to transfer information from similar diseases when prioritizing genes according to the disease of interest.

As discussed in Section 3.3, we assume diseases are structured as a graph $H = (C, \mathbf{\Psi})$, where the $\mathbf{\Psi} = \psi_{kr}|_{r,k=1}^{|C|}$ is the disease similarity matrix, and $\psi_{kr} := \psi_X(k, r) \geq 0$, with X denoting one of the disease similarity measures described in the same section.

To appropriately define the third coordinate of our model, we first focus on which characteristics such coordinate must possess. Intuitively, we expect that the third coordinate $x_{i3}^{(k)}$ for gene i and disease k, embedding information from diseases related to k, respects the following simple properties:

(i) $x_{i3}^{(k)}$ is larger when the gene i is already associated with another disease similar to k, matching the idea that when a gene is known being involved in a disease sharing different profiles (toxicological, genetic, phenotypic

and so on) with another disease, that gene is more likely to have a role even in the latter disease;

(ii) the larger the number of diseases similar to k the gene i is associated with, the larger $x_{i3}^{(k)}$;

(iii) when the gene i has no associations with other diseases, $x_{i3}^{(k)}$ assumes the lowest value;

(iv) an association of gene i with another disease sharing no commonalities with k should be irrelevant for the $x_{i3}^{(k)}$ value.

Before providing our definition of $x_{i3}^{(k)}$, let us introduce some simple definitions. By $\boldsymbol{M}_{i\cdot}$ and $\boldsymbol{M}_{\cdot r}$ we denote respectively the i-th row and the r-th column of a matrix $\boldsymbol{M}$. Moreover, let $L_+^{(k)}$ be the set of positive vertices for disease $k \in C$, and $\boldsymbol{\chi}^{(+,k)}$ its characteristic vector, i.e. $\chi_i^{(+,k)} = 1$ if $i \in L_+^{(k)}$ and $\chi_i^{(+,k)} = 0$ if $i \in V \setminus L_+^{(k)}$. Finally, the matrix containing the characteristic vectors of diseases C is denoted by $\boldsymbol{\mathcal{L}} = (\boldsymbol{\chi}^{(+,1)}, \ldots, \boldsymbol{\chi}^{(+,m)})$.

The *disease semantic relatedness* coordinate for a gene $i \in V$ when predicting a disease $k \in C$ is the following:

$$x_{i3}^{(k)} = \boldsymbol{\mathcal{L}}_{i\cdot} \cdot \boldsymbol{\Psi}_{\cdot k} \tag{12}$$

where ' $\cdot$ ' is the dot product in the Euclidean space $\mathbb{R}^m$. $\boldsymbol{\mathcal{L}}_{i\cdot}$ is thereby the positive labeling of gene i for all considered diseases, and $\boldsymbol{\Psi}_{\cdot k}$ is the vector of similarities between the disease k and the other diseases.

It is straightforward showing that the four prerequisites (i), (ii), (iii), (iv) are satisfied by the definition (12) (the proof can be easily obtained by the definitions of dot product and of the vectors $\boldsymbol{\mathcal{L}}_{i\cdot}$ and $\boldsymbol{\Psi}_{\cdot k}$). Importantly, in order not to introduce bias in the procedure, when computing x_{i3} we exclude descendants r of disease k in the MeSH hierarchy (that is $k \in \text{anc}(r)$), since descendant terms are subsets of their ancestors (see Section 3.3). Furthermore, it may happen, mainly when diseases have few positive genes, that two different but related diseases have exactly the same positive genes (we also experimentally found it in data described in Section 2). Although this would not be a bias but an advantage of our method, in order to have an idea of the performance when all diseases posses different labelings, in our validation experiments we also discarded diseases having exactly the same positive genes of disease k. Hence, the performance of Gene2DisCo shown in Sections 4.1 and 4.2 are suboptimal. Instead, to provide predictions as much reliable as possible, we exploited all diseases when inferring novel candidate disease-genes (see Section 4.3).

To investigate whether this novel feature is really able to improve the discriminative power of our generalized linear models, in Figure 4 we graphically visualize points with coordinates (x_1, x_2, x_3) as done in Figure 3 (a-b) for the two MeSH diseases taken as example, and then draw the best linear regression model separating positive and negative points. We omit by purpose the superscript (k) to simplify the notation.

Interestingly, the added coordinate x_3 (computed using the Jaccard similarity) moves the majority of positive points above negative points: indeed, most

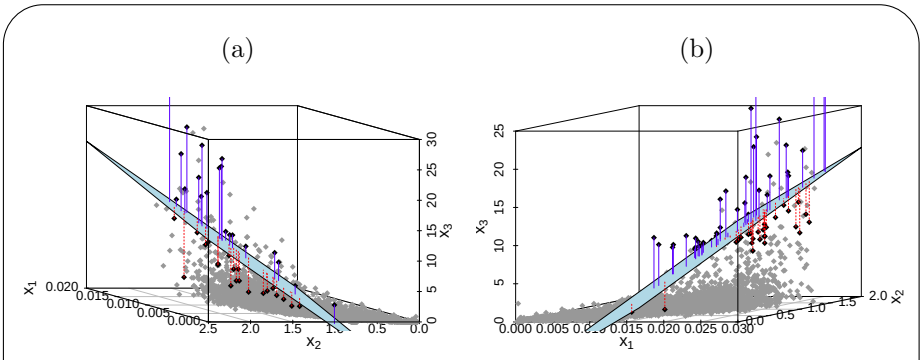

Figure 4: Linear fit of the positive (black) and negative (grey) points shown in Figure 2 represented in the three-dimensional space having coordinates (x_1, x_2, x_3) (a-b). Vertical continuous (blue) and dotted (red) lines represent respectively positive and negative residuals of positive points.

negative points still lie nearby the (x_1, x_2) plane. The learned model now correctly classifies almost all positive points (especially for the example (b)), and positives located below the separator plane however lie not far from it. This is better shown by the residuals of positive points (vertical lines): they are positive in the majority of cases (which means a positive point correctly classified), and the negative residual (corresponding to misclassified positives) have in average a smaller magnitude than the positive residuals. We remark that the two MeSH disease terms shown in this example are representative of a general behavior observed even for the other diseases. As assessed in the next section, the effectiveness of this novel feature is also experimentally confirmed.

Overall, the pseudocode of Gene2DisCo algorithm is shown in Figure 5. Lines 1-4 computes the node projection onto $\mathbb{R}^3$, whose time complexity is $\mathcal{O}(n^2 + nm) = \mathcal{O}(n^2)$ (which becomes $\mathcal{O}(n)$ when $\boldsymbol{W}$ is sparse, like in our context), whereas line 6 performs the fuzzy clustering, taking time $\mathcal{O}(I_1 h^2 |L_+^{(k)}|)$, where I_1 the number of iterations in one run of the clustering algorithm. With complexity $\mathcal{O}(n)$ lines 9-12 calculate the instance weights for the misclassification costs to be used in the GLM training, which is performed at line 13. The cost of training GLM model training depends on the number of iterations I_2 to convergence of the iterative reweighted least square algorithm [67], each one with complexity (considering the weighted least square solution) $\mathcal{O}(m^3 + nm^2 + n^2 m + nm) = \mathcal{O}(n^2 m)$, since $n > m$. In summary, the Gene2DisCo complexity is $\mathcal{O}(n^2 + I_1 h^2 |L_+^{(k)}| + I_2 n^2 m) = \mathcal{O}(I_2 n^2)$, since $m = 3$ and h limited by the number of positives. The source code of Gene2DisCo can be found at `http://frasca.di.unimi.it/data/gene2disco/`.

Figure 5: ***Gene2DisCO* algorithm**.

`Input`:
- gene set V, with $n = |V|$
- disease set C, with $m = |C|$
- $k \in C$ disease to be predicted
- bipartition (U, L) of V
- bipartition ($L_+^{(k)}$, $L_-^{(k)}$) of L
- $n \times n$ gene similarity matrix $\boldsymbol{W}$
- $m \times m$ disease similarity matrix $\boldsymbol{\Psi}$
- $n \times m$ characteristic matrix of positive sets $\boldsymbol{\mathcal{L}}$
- distance d in $\mathbb{R}^3$

`begin algorithm`

01: `for each` $i \in V$ `do`

02: $x_{i1} \leftarrow \sum_{j \in L_+^{(k)}} W_{ij}$; $x_{i2} \leftarrow \sum_{j \in L_-^{(k)}} W_{ij}$; $x_{i3} \leftarrow \sum_{r=1}^{m} \mathcal{L}_{ir} \Psi_{rk}$

03: $\Delta_i := (x_{i1}, x_{i2}, x_{i3})$

04: `end for`

05: $P_+^{(k)} := \{\Delta_i | i \in L_+^{(k)}\}$; $P_-^{(k)} := \{\Delta_i | i \in L_-^{(k)}\}$

06: $\sigma \leftarrow \texttt{FCM}\big(P_+^{(k)}, \alpha, d\big)$

07: $z \leftarrow 0.5$; $\tau \leftarrow \texttt{quantile}(\sigma(P_-^{(k)}), z)$

08: $\bar{L}_-^{(k)} := \{i \in L_- | \sigma(\Delta_i) > \tau\}$; $\bar{P}_-^{(k)} = \{\Delta_i | i \in \bar{L}_-\}$

09: `for each` i in $\{L_+^{(k)} \cup \bar{L}_-^{(k)}\}$ `do`

10: `if` i in $L_+^{(k)}$ `then` $\omega_i \leftarrow x_{i1} / \sum_{j \in L_+^{(k)}} x_{j1}$

11: `else` $\omega_i \leftarrow (1 - \sigma(\Delta_i)) / \sum_{j \in \bar{L}_-^{(k)}} (1 - \sigma(\Delta_j))$

12: `end for`

13: $\phi(x_1, x_2, x_3) \leftarrow \texttt{GLM}(P_+^{(k)}, \bar{P}_-^{(k)}, \boldsymbol{\omega}, \texttt{link} = \texttt{Cloglog})$

`end algorithm`

`Output`: the ranking function ϕ and the predicted scores $\phi(\Delta_i),\ i \in U$.

4. Results and discussion

Gene2DisCo has been experimentally validated by following the benchmark setting proposed in [47], thus allowing its fair comparison with the benchmark methodologies (described in the next section). Namely, such setting adopts the k-fold cross-validation (CV) procedure ($k = 5$) to assess the generalization abilities of the compared methods, and the *Area Under the ROC Curve* (AUC) and the *Precision* at different *Recall* levels (PXR) to measure the corresponding performance. Furthermore, as done by authors of *NWGP* on the same data, we also computed for our method the Area Under the Precision-Recall Curve (AUPRC), being AUPRC more informative than AUC on unbalanced settings [68].

The benchmark methods adopted in [47] are the following: *GBA*, family of algorithms relying upon the guilt-by-association rule, which allows making

predictions exploiting the interacting genes, under the assumption that interacting genes are likely to share similar functions [69, 70]; *RW*, random walks algorithm [71]; *RWR*, random walks with restart which takes into account that after many steps the walker may forget the prior information coded in the initial probability vector, and accordingly the algorithm allows the walker to restart from its initial condition with a given probability θ (free parameter), or to move another random walk step with probability $1 - \theta$; *kernelized score functions*, extending the similarity to non neighboring nodes by adopting a suitable kernel matrix [72]. The score for each gene i with regard to a given disease r is defined according to a suitable distance $d(i, V_r)$ between i and the subset V_r of positive genes for r. By varying the definition of $d(i, V_r)$ authors obtained different scoring methods, among which the top performing was S_{AV}: $d(i, V_r)$ is defined as the average distance between the images in the corresponding Hilbert space of i and V_r. The kernel used is $\boldsymbol{K}^t$ obtained by a t-step ($t = 1,\ 2,\ \ldots$) random walk, where $\boldsymbol{K} = \gamma \boldsymbol{I} + \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{W} \boldsymbol{D}^{-\frac{1}{2}}$, $\boldsymbol{I}$ is the $n \times n$ identity matrix, $\gamma > 0$, and $\boldsymbol{D}$ is a diagonal matrix whose diagonal elements are the sums of the corresponding rows in $\boldsymbol{W}$.

As baseline comparison, in this paper we report just the performance of the top benchmark method (S_{AV} $t = 5$). Furthermore, we compare our method also with the *NWGP* algorithm.

4.1. Evaluation on benchmark data

We dedicate this section to the comparison of our method with benchmark methods described above, and to the evaluation of the impact the different disease similarity measures described in Section 3.3 have on the Gene2DisCo performance. To this end, we have constructed other six similarity measures by integrating the considered measures through the unweighted sum (US) integration and the Max integration, defined as follows. Given s disease similarity matrices $\boldsymbol{\Psi}^{(1)},\ \ldots, \boldsymbol{\Psi}^{(s)}$, the unweighted sum integration is

$$\boldsymbol{\Psi} = \sum_{k=1}^{s} \boldsymbol{\Psi}^{(k)} / s \ ,$$

whereas the *Max* integration is obtained by setting each entry of the integrated matrix as the maximum value of that entry in the single matrices:

$$\psi_{ij} = \max_{k \in \{1,\ldots,s\}} \psi_{ij}^{(k)},\ \forall i, j \in \{1, \ldots, |C|\}$$

We employed both schemes in three different integrations: (*ALL*) integration of all considered measures; (*InfoSim*) integration of information-based measures only; (*PathSim*) integration of solely path-based measures. Figure 6 shows the corresponding results. To facilitate distinguishability among methods, bar colors have been grouped according to the type of diseases semantic similarity: path-based, information-based, US and Max similarity measures are respectively in blue, turquoise, green and brown scale colors.

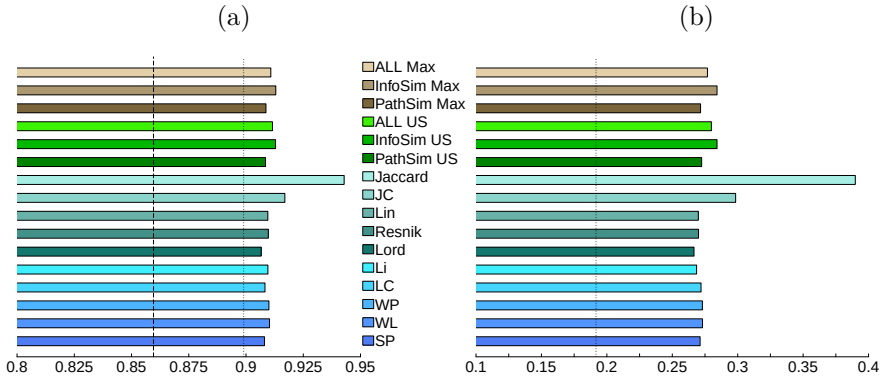

Figure 6: AUC (a) and AUPRC (b) averaged across diseases for Gene2DisCo method when adopting different disease semantic similarity measures. '*US*' and '*Max*' refer to the integrated measures according to the unweighted sum and the Max integration, respectively. 'ALL' denotes the integration of both path- and information-based measures. Vertical dotted lines correspond to *NWGP* results, whereas the vertical dashed line represents the best benchmark results (S_{AV} $t = 5$).

Firstly, results have an analogous trend for both AUC and AUPRC measures, with most measures showing similar performance. The only exception is the Jaccard similarity, which largely achieves the best results (Wilcoxon signed rank test, *p-value* < 0.001) in both performance metrics, especially in terms of AUPRC, which is more relevant in this context (rare positives). Such an advantage for the Jaccard similarity also confirms results obtained in [73]. Gene2DisCo outperforms *NWGP* in both AUC and AUPRC, with any disease semantic similarity metric adopted, mainly in terms of AUPRC, showing the noticeable benefit of transferring information from similar diseases. Indeed, the improvement when using the Jaccard similarity is impressive: 0.1917 for *NWGP* and 0.3899 for Gene2DisCo, more than the double. By reminding AUPRC measures the area under the Precision and Recall curve, this means in particular that the ability of our method in classifying positives is considerably improved with reference to *NWGP*: in the gene prioritization context, this corresponds to a better ranking of disease-genes, which is exactly the final aim of gene prioritization. Even S_{AV} $t = 5$ is largely outperformed by Gene2DisCo in terms of AUC (we remind that in the benchmark setting, AUPRC was not computed). This behaviour is confirmed by the PXR results in Figure 7: all Gene2DisCo variants lie above both S_{AV} $t = 5$ and *NWGP*, with larger improvements for smaller Recall values, which are more important in context with scarcity of positives. Even in this case, the Jaccard variant performs best among Gene2DisCo variants, for all the Recall values considered.

Finally, Gene2DisCo takes around 7 seconds to perform a 5-fold CV procedure for a single MeSH disease term on an Intel i7-860 CPU 2.80 GHz machine

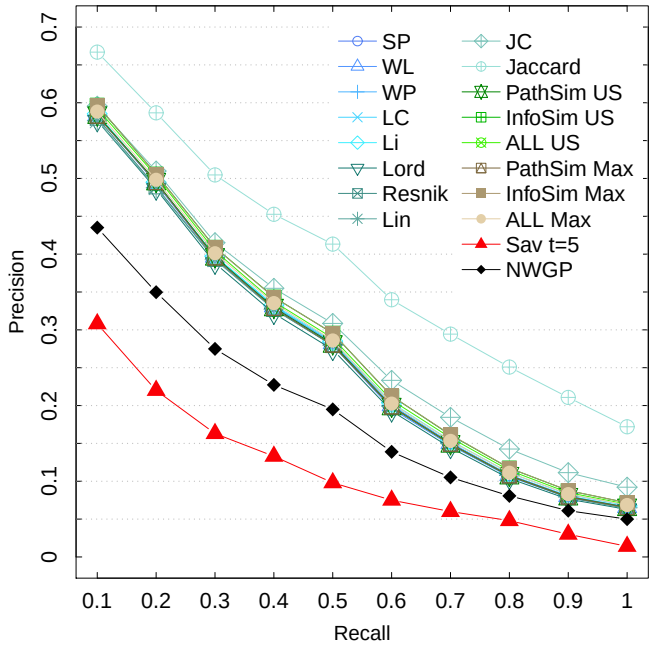

Figure 7: PXR values of the different variants of Gene2DisCo, *NWGP* and of the original top benchmark method S_{AV} $t = 5$.

with 16 GB of RAM, including the computation of Jaccard similarity on the training data. Hence, passing from 2- to 3-dimensional space just slightly increased the overall computational complexity of the GLM training, considering that *NWGP* in the same setting, but without the computation of the Jaccard similarity, takes around 5 seconds.

4.2. Results on refined benchmark data

Although our aim on the enriched data set is to infer novel candidate disease-genes, since we have already validated our method in the Section 4.1, to facilitate possible comparisons with other methodologies, in Table 1 we report the average results obtained through a 5-fold cross validation in predicting the 470 selected MeSH disease terms in the refined benchmark data described in Section 2.2. It is interesting to observe that the quality of Gene2DisCo predictions, with reference to results shown in Section 4.1, is maintained even in this setting with more genes and higher complexity (stronger labeling imbalance). Moreover, to better analyze the behaviour of Gene2DisCo, we distinguished average results across the two categories of diseases with 5-200 and 2-4 disease-genes. Interestingly, the performance in the former category of diseases, those having less available information, is not affected by a large overall decay: an average AUPRC value 0.268 is remarkable for diseases with very rare positives.

AUC	AUPRC	PXR									
		Precision									
		Rec = 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
					2-200						
0.943	0.344	0.562	0.495	0.427	0.387	0.357	0.313	0.281	0.255	0.215	0.176
					2-4						
0.725	0.268	0.321	0.321	0.321	0.308	0.308	0.266	0.265	0.265	0.265	0.265
					5-200						
0.959	0.350	0.580	0.508	0.435	0.393	0.361	0.316	0.283	0.255	0.211	0.169

Table 1: Gene2DisCo performance on the enriched data set averaged across diseases with 2-200, 2-4 and 5-200 known disease-genes. 'Rec' stands for recall.

Level									
1	2	3	4	5	6	7	8	9	10
0	8	80	183	242	206	92	23	5	3

Table 2: Level distribution in the MeSH taxonomy for the selected diseases.

4.3. Providing novel putative gene-disease associations

This section is dedicated to provide predictions for novel candidate gene-disease associations. Since our method can be also exploited to predict diseases with one or zero known disease-genes, we included in our setting also such diseases, in addition to those described in Section 2.2. We did not considered these diseases in the previous section because they cannot be used in cross validation procedures, since positives (disease-genes) are not enough. Specifically, we selected: 1) diseases with at most 200 genes associated with; 2) diseases without known gene associations but sharing non null disease commonalities with at least one disease with positives (348). We adopted the Jaccard similarity (which performed best) for the diseases with positives, and the sum integration of path-based similarity measures (Jaccard is not applicable, see Section 3.3) for diseases without positives.

We obtained a set of 842 MeSH disease terms, spanning all the different levels in the hierarchy except for level 1 (root nodes, we remind that the level is the number of nodes on the path corresponding to the maximum distance from a root node). The empirical distribution across node levels is reported in Table 2. Furthermore, the selected diseases cover different MeSH trees, whose roots are shown in Table 3, with the corresponding description. The total number of diseases in that table is larger than 842 because some MeSH headings fall in more than one tree.

To infer the gene ranking for MeSH disease terms without annotated genes,

Tree root	Diseases	Root description
C02	34	Virus Diseases
C03	40	Parasitic Diseases
C04	103	Neoplasms
C05	103	Musculoskeletal Diseases
C06	52	Digestive System Diseases
C07	26	Stomatognathic Diseases
C08	36	Respiratory Tract Diseases
C10	193	Nervous System Diseases
C11	48	Eye Diseases
C12	38	Male Urogenital Diseases
C13	54	Female Urogenital Diseases and Pregnancy Complications
C14	59	Cardiovascular Diseases
C15	69	Hemic and Lymphatic Diseases
C16	326	Congenital, Hereditary, and Neonatal Diseases and Abnormalities
C17	94	Skin and Connective Tissue Diseases
C18	142	Nutritional and Metabolic Diseases
C19	39	Endocrine System Diseases
C20	38	Immune System Diseases
C21	1	Disorders of Environmental Origin
C22	13	Animal Diseases
C23	86	Pathological Conditions, Signs and Symptoms
C24	1	Occupational Diseases
C25	6	Chemically-Induced Disorders
C26	27	Wounds and Injuries
F03	24	Mental Disorders

Table 3: Tree root categories of the selected MeSH disease terms.

we could not learn the GLM model because there are no positives; accordingly we used the inter-disease score in formula (12) to determine an association score for gene i and disease k. The obtained score matrix $\boldsymbol{A} = \boldsymbol{\mathcal{L}} \cdot \boldsymbol{\Psi}$ can be downloaded at `https://frasca.di.umimi.it/gene2disco/`. In order to facilitate interpreting such predictions, we selected for every disease the three top-ranked genes. Since the inferred score may depend on the number of diseases a gene is already associated with (see formula 12), we added a further processing step for determining the desired list:

1. For every $k \in \{1, 2, \ldots, m\}$, sort each row $\boldsymbol{A}_{i\cdot}$ in decreasing order
2. Determine the rank $r_i^{(k)}$ of disease k in the obtained vector
3. Select the genes corresponding to the three highest ranks in the vector $\boldsymbol{r}^k$, that is the genes for which disease k was ranked atop the other diseases

For diseases with known associations, we adopted the GLM prediction as discussed in Section 3.4.3 to determine the top three genes for every disease. In order to provide the most reliable predictions, a more intensive procedure has been adopted. Namely, we used all the positive genes available to train the model, and randomly partitioned negative genes in ten subsets, using nine of them together with positives for training a model to predict the remaining subset. The procedure has been iterated ten times, one for every subset of negatives. All inferred predictions and selected candidate disease-genes can be found at `https://frasca.di.umimi.it/gene2disco/`.

As further validation, we ran multiple independent runs of the cross validation procedure and computed the inferred putative associations confirmed in each run. To investigate the most specific descriptions and more reliable predictions, we considered the MeSH disease terms with the highest performance in all the runs and with at most 10 positive genes. Moreover, in the obtained associations we further filter out those involving genes with high node degree in the network, being node degree a proxy for gene multifunctionality and thus leading to more generic associations [74]. The suggested candidate disease-genes are shown in Table 4.

Disease name	**MeSH ID**	**Genes**	**Candidate genes**
Angioid Streaks	D000793	2	ABCB1
Central Serous Chorioretinopathy	D056833	10	IL1B
Psychoses, Alcoholic	D011604	2	CYP3A4
Herpes Zoster Ophthalmicus	D006563	10	ACMSD,AFMID
Herpes Zoster Oticus	D006563	10	ACMSD,AFMID
Tennis Elbow	D013716	8	LGALS13
Neuroaspergillosis	D020953	7	GSTA5
Olfactory Nerve Injuries	D061219	10	IL1F7
Diffuse alopecia, Felty Syndrome	C531609, D005258	3	CYP2D6,CASR

Table 4: Genes and MeSH headings selected as candidate novel gene-disease associations. The column 'Genes' contains the number of known disease-genes (CTD database, update 04.17).

Interestingly, the protein coding ABCB1 (ATP binding cassette subfamily B member 1) and ABCC6 (ATP binding cassette subfamily C member 6) belong to the same superfamily of ATP-binding cassette (ABC), and multiple studies have already associated ABCC6 with *Angioid Streaks* [75, 76, 77]. The protein coding gene ILB1 (interleukin 1 beta) is a member of the interleukin family, to whom belong the interleukins IL12B IL5RA, IL3, IL5 having curated associations with *Central Serous Chorioretinopathy* at CTD database (04.17). Gene CYP3A4, encoding a member of the cytochrome P450 superfamily of enzymes, is already associated with *Delirium* [78], *Depressive Disorder* [79], and *Bipo-*

lar Disorder [80], diseases sharing therapeutic chemical profiles with *Psychoses, Alcoholic* (CTD *disease comp*, update 04.17). Moreover, gene IL1F7 is a member of the same family of two genes, IL1A and IL1B, actually associated with *Olfactory Nerve Injuries* [81]. Finally, gene CASR (calcium sensing receptor) is involved in *Hyperparathyroidism, Primary* disease, which shares chemical interaction profiles (via *Lithium Carbonate*, CTD *disease comp*, update 04.17) with *Diffuse alopecia* disease.

5. Conclusions

This work showed that disease relatedness is a fundamental feature of the gene prioritization (GP) process, and that GP methods must exploit commonalities among diseases to achieve more reliable predictions. We have developed a network-based approach, named Gene2DisCo, for prioritizing genes associated with a given disease, whose main contributions are the following: (i) the construction of a framework aware of several issues characterizing the gene prioritization process, including the need of integrating different data sources, the rarity of known diseases-genes, and the existence of common profiles among different diseases; (ii) the derivation of a method based on generalized linear model (GLM) embedding disease pairwise similarities; (iii) the extensive comparison of several state-of-the-art semantic measures to compute semantic similarity among diseases. With this novel technique, our method performed much better than a set of state-of-the-art methodologies for disease-gene prioritization on a benchmark dataset and than an existing technique based on GLM using exactly the same input information as Gene2DisCo, but neglecting information from similar diseases. Unlike existing approaches, a main feature of Gene2DisCo is the capability of inferring gene prioritization lists even for diseases without known causative genes. This study also made available a novel benchmark dataset for future comparison providing gene-gene connections extended to the whole human genome, and recent genes-disease associations retrieved from the Medical Subjects Headings (MeSH) ontology. Finally, exploiting the ability of Gene2DisCo in enriching the information available for diseases with few or none associations, we provided a set of three top-ranked candidate genes for 842 MeSH diseases with 0 up to 200 known gene associations.

ACKNOWLEDGEMENTS

This work was funded by the Department of Computer Science of the University of Milan under the grant "Piano di sostegno alla ricerca 2015/2017" (grant number 19246), project title *Graph-based methodologies for the automated inference in bio-medical ontologies*.

References

[1] S. M. Goldman, Environmental toxins and parkinson's disease, Annual Review of Pharmacology and Toxicology 54 (1) (2014) 141–164, pMID: 24050700. doi:10.1146/annurev-pharmtox-011613-135937.

[2] Y. Moreau, L.-C. Tranchevent, Computational tools for prioritizing candidate genes: boosting disease gene discovery, Nat Rev Genet 13 (8) (2012) 523–536. doi:10.1038/nrg3253.

[3] J. Amberger, C. Bocchini, A. Hamosh, A new face and new challenges for online mendelian inheritance in man (omim), Human Mutation 32 (5) (2011) 564–567. doi:10.1002/humu.21466.

[4] J. Fontaine, A.-N. M., Gene Set to Diseases (GS2D): Disease Enrichment Analysis on Human Gene Sets with Literature Data, Genomics and Computational Biology 2 (1) (2016) e33.

[5] M. Krallinger, A. Valencia, L. Hirschman, Linking genes to literature: text mining, information extraction, and retrieval applications for biology, Genome Biology 9 (2) (2008) S8. doi:10.1186/gb-2008-9-s2-s8.

[6] R. Winnenburg, T. Wchter, C. Plake, A. Doms, M. Schroeder, Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies?, Briefings in Bioinformatics 9 (6) (2008) 466–478.

[7] V. G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response., Proceedings of the National Academy of Sciences of the United States of America 98 (9) (2001) 5116–5121. arXiv:1501.01455, doi:10.1073/pnas.091062498.

[8] G. Zheng, B. Freidlin, J. L. Gastwirth, Robust genomic control for association studies, The American Journal of Human Genetics 78 (2) (2006) 350 – 356. doi:http://dx.doi.org/10.1086/500054.

[9] B. Lehne, C. M. Lewis, T. Schlitt, From snps to genes: Disease association at the gene level, PLoS ONE 6 (6) (2011) e20133. doi:10.1371/journal.pone.0020133.

[10] L. Franke, et al., Team: a tool for the integration of expression, and linkage and association maps, European journal of human genetics 12 (8) (2004) 633–638.

[11] T. A. Manolio, Genome-wide association studies and assessment of the risk of disease, New England Journal of Medicine 363 (2) (2010) 166–176. doi:10.1056/NEJMra0905980.

[12] L. Mamanova, A. J. Coffey, et al., Target-enrichment strategies for next-generation sequencing, Nat Meth 7 (2) (2011) 111–118. doi:10.1038/nmeth.1419.

[13] J. de Ligt, M. H. Willemsen, et al., Diagnostic exome sequencing in persons with severe intellectual disability, New England Journal of Medicine 367 (20) (2012) 1921–1929. doi:10.1056/NEJMoa1206524.

[14] D. Smedley, P. N. Robinson, Phenotype-driven strategies for exome prioritization of human mendelian disease genes, Genome Medicine 7 (1) (2015) 1–11. doi:10.1186/s13073-015-0199-2.

[15] D. Brnigen, L.-C. Tranchevent, F. Bonachela-Capdevila, K. Devriendt, B. D. Moor, P. D. Causmaecker, Y. Moreau, An unbiased evaluation of gene prioritization tools, Bioinformatics 28 (23) (2012) 3081–3088. doi:10.1093/bioinformatics/bts581.

[16] T. Ideker, R. Sharan, Protein networks in disease, Genome Research 18 (4) (2008) 644–652. doi:10.1101/gr.071852.107.

[17] S. Navlakha, C. Kingsford, The power of protein interaction networks for associating genes with diseases, Bioinformatics (Oxford, England) 26 (8) (2010) 10571063. doi:10.1093/bioinformatics/btq076.

[18] S. Navlakha, et al., Finding biologically accurate clusterings in hierarchical tree decompositions using the variation of information, in: S. Batzoglou (Ed.), Research in Computational Molecular Biology, Vol. 5541 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2009, pp. 400–417.

[19] O. Vanunu, R. Sharan, A Propagation-based Algorithm for Inferring Gene-Disease Associations, in: Proceedings of the German Conference on Bioinformatics, GCB, September 9-12, Dresden, Germany, 2008.

[20] S. Kohler, S. Bauer, D. Horn, P. N. Robinson, Walking the interactome for prioritization of candidate disease genes, The American Journal of Human Genetics 82 (4) (2008) 949 – 958. doi:10.1016/j.ajhg.2008.02.013.

[21] M. Re, G. Valentini, Cancer module genes ranking using kernelized score functions, BMC Bioinformatics 13 (Suppl 14) (2012) S3. doi:10.1186/1471-2105-13-S14-S3.

[22] T.-P. Nguyen, T.-B. Ho, Detecting disease genes based on semi-supervised learning and protein-protein interaction networks, Artificial Intelligence in Medicine 54 (1) (2012) 63 – 71. doi:10.1016/j.artmed.2011.09.003.

[23] F. Mordelet, J.-P. Vert, Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples, BMC Bioinformatics 12 (1) (2011) 389. doi:10.1186/1471-2105-12-389.

[24] H. Zhou, J. Skolnick, A knowledge-based approach for predicting gene-disease associations, Bioinformatics 32 (18) (2016) 2831. doi:10.1093/bioinformatics/btw358.

[25] T. Yin, X. Wu, W. Tian, GenePANDA-a novel network-based gene prioritizing tool for complex diseases, Scientific Reports 7 (43258). doi:10.1038/srep43258.

[26] L.-C. Tranchevent, et al., Endeavour update: a web resource for gene prioritization in multiple species, Nucleic Acids Research 36 (suppl 2) (2008) W377–W384. doi:10.1093/nar/gkn325.

[27] A. Oellrich, Sanger Mouse Genetics Project, D. Smedley, Linking tissues to phenotypes using gene expression profiles, Database : the journal of biological databases and curation 2014 (2014) bau017. doi:10.1093/database/bau017.

[28] A. Antanaviciute, et al., GeneTIER: prioritization of candidate disease genes using tissue-specific gene expression profiles, Bioinformatics (Oxford, England) 31 (16) (2015) 27282735. doi:10.1093/bioinformatics/btv196.

[29] J. Che, M. Shin, A meta-analysis strategy for gene prioritization using gene expression, snp genotype, and eqtl data, BioMed Research International 2015 (2015) 1–8. doi:doi:10.1155/2015/576349.

[30] A. C. Gavin, et al., Proteome survey reveals modularity of the yeast cell machinery, Nature. 440 (7084) (2006) , 631–636. doi:10.1038/nature04532.

[31] A. Antanaviciute, et al., Ova: integrating molecular and physical phenotype data from multiple biomedical domain ontologies with variant filtering for enhanced variant prioritization, Bioinformatics-doi:10.1093/bioinformatics/btv473.

[32] M. Frasca, et al., UNIPred: unbalance-aware Network Integration and Prediction of protein functions, Journal of Computational Biology 22 (12) (2015) 1057–1074.

[33] S. Mathur, D. Dinakarpandian, Finding disease similarity based on implicit semantic similarity, Journal of Biomedical Informatics 45 (2) (2012) 363 – 371. doi:http://dx.doi.org/10.1016/j.jbi.2011.11.017.

[34] A. Oellrich, et al., Improving disease gene prioritization by comparing the semantic similarity of phenotypes in mice with those of human diseases., PloS one 7 (6) (2012) e38937+. doi:10.1371/journal.pone.0038937.

[35] M. Vidal, M. E. Cusick, A.-L. Barabsi, Interactome networks and human disease, Cell 144 (6) (2011) 986 – 998. doi:http://dx.doi.org/10.1016/j.cell.2011.02.016.

[36] A. P. Davis, et al., The Comparative Toxicogenomics Database: update 2013, Nucleic Acids Research 41 (D1) (2013) D1104–D1114. doi:10.1093/nar/gks994.

[37] S. Craft, et al., Intranasal insulin therapy for alzheimer disease and amnestic mild cognitive impairment: a pilot clinical trial, Archives of neurology 69 (1) (2012) 29–38.

[38] C. Elkan, The foundations of cost-sensitive learning, in: In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, 2001, pp. 973–978.

[39] C. X. Ling, V. S. Sheng, Cost-sensitive Learning and the Class Imbalanced Problem, 2007, encyclopedia of Machine Learning.

[40] M. Frasca, Automated gene function prediction through gene multifunctionality in biological networks, Neurocomputing 162 (0) (2015) 48 – 56. doi:10.1016/j.neucom.2015.04.007.

[41] M. Frasca, S. Bassis, G. Valentini, Learning node labels with multi-category Hopfield networks, Neural Computing and Applications (2015) 1–16doi:10.1007/s00521-015-1965-1.

[42] J. , H. C. Wick, D. E. Kee, K. Noto, J. L. Maron, D. K. Slonim, Finding novel molecular connections between developmental processes and disease, PLOS Computational Biology 10 (5) (2014) 1–12. doi:10.1371/journal.pcbi.1003578.

[43] M. Frasca, S. Bassis, Gene-Disease Prioritization Through Cost-Sensitive Graph-Based Methodologies, Vol. 9656 of Lecture Notes in Computer Science, Springer International Publishing, Cham, 2016, pp. 739–751. doi:10.1007/978-3-319-31744-1_64.

[44] M. Frasca, D. Malchiodi, Exploiting negative sample selection for prioritizing candidate disease genes, Genomics and Computational Biology 3 (3) (2017) 47. doi:10.18547/gcb.2017.vol3.iss3.e47.

[45] D. Szklarczyk, et al., String v10: protein-protein interaction networks, integrated over the tree of life, Nucleic Acids Research 43 (D1) (2015) D447–D452. doi:10.1093/nar/gku1003.

[46] A. P. Davis, et al., Comparative toxicogenomics database: a knowledgebase and discovery tool for chemical-gene-disease networks, Nucleic acids research 37 (Database issue) (2009) D78692. doi:10.1093/nar/gkn580.

[47] G. Valentini, et al., An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods, Artificial Intelligence in Medicine 61 (2) (2014) 63 – 78. doi:http://dx.doi.org/10.1016/j.artmed.2014.03.003.

[48] S. Nelson, M. Schopen, A. Savage, J. Schulman, N. Arluk, The mesh translation maintenance system: Structure, interface design, and implementation, in: Proceedings of the 11th World Congress on Medical Informatics, IOS Press, 2004, pp. 67–69.

[49] M. Ashburner, et al., Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium., Nature genetics 25 (1) (2000) 25–29.

[50] G. Wu, X. Feng, L. Stein, A human functional protein interaction network and its application to cancer data analysis, Genome Biology 11 (5) (2010) 1–23. doi:10.1186/gb-2010-11-5-r53.

[51] I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, E. M. Marcotte, Prioritizing candidate disease genes by network-based boosting of genome-wide association data, Genome research 21 (7) (2011) 11091121. doi:10.1101/gr.118992.110.

[52] E. Segal, N. Friedman, D. Koller, A. Regev, A module map showing conditional activity of expression modules in cancer, Nature Genetics 36 (3) (2004) 1090–1098.

[53] A. Chatr-aryamontri, B.-J. Breitkreutz, S. Heinicke, L. Boucher, et al., The biogrid interaction database: 2013 update., Nucleic Acids Research 41 (Database-Issue) (2013) 816–823.

[54] P. Resnik, Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language, Journal of Artificial Intelligence Research 11 (1999) 95–130.

[55] S. Kohler, et al., The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data, Nucleic Acids Research 42 (D1) (2014) D966–D974. doi:10.1093/nar/gkt1026.

[56] H. Bulskov, R. Knappe, T. Andreasen, On Measuring Similarity for Conceptual Querying, in: FQAS '02: Proceedings of the 5th International Conference on Flexible Query Answering Systems, Springer-Verlag, London, UK, 2002, pp. 100–111.

[57] R. Richardson, A. Smeaton, J. Murphy, Using wordnet as a knowledge base for measuring semantic similarity between words, in: Proc. AICS Conference, Trinity College, Dublin, 1994.

[58] Z. Wu, M. Palmer, Verbs semantics and lexical selection, in: Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics, ACL '94, Association for Computational Linguistics, Stroudsburg, PA, USA, 1994, pp. 133–138. doi:10.3115/981732.981751.

[59] C. Leacock, M. Chodorow, Filling in a sparse training space for word sense identification, ms., March. (1994).

[60] Y. Li, , et al., An approach for measuring semantic similarity between words using multiple information sources, IEEE Trans. on Knowl. and Data Eng. 15 (4) (2003) 871–882. doi:10.1109/TKDE.2003.1209005.

[61] P. Lord, Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation, Bioinformatics 19 (10) (2003) 1275–1283. doi:10.1093/bioinformatics/btg153.

[62] D. Lin, An information-theoretic definition of similarity, in: Proceedings of the Fifteenth International Conference on Machine Learning, ICML '98, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998, pp. 296–304.

[63] J. J. Jiang, D. W. Conrath, Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, in: International Conference Research on Computational Linguistics (ROCLING X), 1997, pp. 9008+.

[64] P. Jaccard, Lois de distribution florale dans la zone alpine, imprimerie Corbaz & Comp. (1902). doi:10.5169/seals-266762.

[65] M. Frasca, D. Malchiodi, Selection of negative examples for node label prediction through fuzzy clustering techniques, in: Advances in Neural Networks: Computational Intelligence for ICT, Springer International Publishing, Cham, 2016, pp. 67–76. doi:10.1007/978-3-319-33747-0_7.

[66] J. C. Bezdek, R. Ehrlich, W. Full, Fcm: The fuzzy c-means clustering algorithm, Computers & Geosciences 10 (2) (1984) 191–203.

[67] J. A. Barreto, C. S. Burrus, Iterative reweighted least squares and the design of two-dimensional FIR digital filters, in: Proceedings 1994 International Conference on Image Processing, Austin, Texas, USA, November 13-16, 1994, 1994, pp. 775–779. doi:10.1109/ICIP.1994.413420.

[68] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets, PLoS ONE 10 (3) (2015) 1–21. doi:10.1371/journal.pone.0118432.

[69] E. Marcotte, M. Pellegrini, M. Thompson, T. Yeates, D. Eisenberg, A combined algorithm for genome-wide prediction of protein function, Nature 402 (1999) 83–86.

[70] B. Schwikowski, P. Uetz, S. Fields, A network of protein-protein interactions in yeast., Nature biotechnology 18 (12) (2000) 1257–1261.

[71] L. Lovász, Random walks on graphs: A survey, in: D. Miklós, V. T. Sós, T. Szőnyi (Eds.), Combinatorics, Paul Erdős is Eighty, Vol. 2, János Bolyai Mathematical Society, Budapest, 1996, pp. 353–398.

[72] G. Valentini, G. Armano, M. Frasca, J. Lin, M. Mesiti, M. Re, RANKS: a flexible tool for node label ranking and classification in biological networks, Bioinformatics 32 (18) (2016) 2872–2874. doi:10.1093/bioinformatics/btw235.

[73] M. Frasca, N. Cesa-Bianchi, Multitask protein function prediction through task dissimilarity, IEEE/ACM Transactions on Computational Biology and Bioinformatics PP (99) (2017) in press. doi:10.1109/TCBB.2017.2684127.

[74] J. Gillis, P. Pavlidis, The Impact of Multifunctional Genes on "Guilt by Association" Analysis, PLoS ONE 6 (2) (2011) e17258+.

[75] Y. Nitschke, G. Baujat, U. Botschen, et al., Generalized arterial calcification of infancy and pseudoxanthoma elasticum can be caused by mutations in either {ENPP1} or {ABCC6}, The American Journal of Human Genetics 90 (1) (2012) 25 – 39. doi:10.1016/j.ajhg.2011.11.020.

[76] R. P. Finger, P. C. Issa, M. S. Ladewig, C. Gtting, C. Szliska, H. P. Scholl, F. G. Holz, Pseudoxanthoma elasticum: Genetics, clinical manifestations and therapeutic approaches, Survey of Ophthalmology 54 (2) (2009) 272 – 285. doi:10.1016/j.survophthal.2008.12.006.

[77] N. Sato, T. Nakayama, Y. Mizutani, M. Yuzawa, Novel mutations of {ABCC6} gene in japanese patients with angioid streaks, Biochemical and Biophysical Research Communications 380 (3) (2009) 548 – 553. doi:10.1016/j.bbrc.2009.01.117.

[78] C.-C. Huang, I.-H. Wei, Unexpected interaction between quetiapine and valproate in patients with bipolar disorder, General Hospital Psychiatry 32 (4) (2010) 446.e1 – 446.e2. doi:10.1016/j.genhosppsych.2009.06.005.

[79] P. Kumar, H. Kalonia, A. Kumar, Novel protective mechanisms of antidepressants against 3-nitropropionic acid induced huntingtons-like symptoms: a comparative study, Journal of Psychopharmacology 25 (10) (2011) 1399–1411, pMID: 20305041. doi:10.1177/0269881110364269.

[80] A. H. Sulaiman, M. A. Said, M. H. Habil, R. Rashid, A. Siddiq, N. C. Guan, M. Midin, N. R. N. Jaafar, H. Sidi, S. Das, The risk and associated factors of methamphetamine psychosis in methamphetamine-dependent patients in malaysia, Comprehensive Psychiatry 55, Supplement 1 (2014) S89 – S94. doi:10.1016/j.comppsych.2013.01.003.

[81] K. N. Corps, Z. Islam, J. J. Pestka, J. R. Harkema, Neurotoxic, inflammatory, and mucosecretory responses in the nasal airways of mice repeatedly exposed to the macrocyclic trichothecene mycotoxin roridin a, Toxicologic Pathology 38 (3) (2010) 429–451. doi:10.1177/0192623310364026.