7-2018

Disease gene classification with metagraph representations

Sezin KIRCALI ATA
Nanyang Technological University

Yuan FANG
Singapore Management University, yfang@smu.edu.sg

Min WU
Institute for Infocomm Research

Xiao-Li LI
Institute for Infocomm Research

Xiaokui XIAO
National University of Singapore

Disease Gene Classification with Metagraph Representations

Abstract

Protein-protein interaction (PPI) networks play an important role in studying the functional roles of proteins, including their association with diseases. However, protein interaction networks are not sufficient without the support of additional biological knowledge for proteins such as their molecular functions and biological processes. To complement and enrich PPI networks, we propose to exploit biological properties of individual proteins. More specifically, we integrate *keywords* describing protein properties into the PPI network, and construct a novel *PPI-Keywords* (PPIK) network consisting of both proteins and keywords as two different types of nodes. As disease proteins tend to have a similar topological characteristics on the PPIK network, we further propose to represent proteins with *metagraphs*. Different from a traditional network motif or subgraph, a metagraph can capture a particular topological arrangement involving the interactions/associations between both proteins and *keywords*. Based on the novel metagraph representations for proteins, we further build classifiers for disease protein classification through supervised learning. Our experiments on three different PPI databases demonstrate that the proposed method consistently improves disease protein prediction across various classifiers, by 15.3% in AUC on average. It outperforms the baselines including the diffusion-based methods (e.g., RWR) and the module-based methods by 13.8–32.9% for overall disease protein prediction. For predicting breast cancer genes, it outperforms RWR, PRINCE and the module-based baselines by 6.6–14.2%. Finally, our predictions also turn out to have better correlations with literature findings from PubMed.

Keywords: protein-protein interaction, Uniprot keywords, metagraph, protein representations, disease protein prediction

1. Introduction

Studying disease-causing genes and their protein products is critical to the diagnosis and treatment of serious diseases such as cancer and diabetes. Despite recent advances in identifying the functions of genes and proteins [1, 2, 3, 4, 5, 6], it still remains a challenging research issue to understand their interactions and pathways in the context of many diseases.

In order to decipher how proteins work together, *protein-protein interaction* (PPI) networks [7, 8, 9] have been widely exploited. Several studies [10, 11] have demonstrated that the locality of a protein in a PPI network is not random. Rather, proteins with the same phenotype or function tend to exhibit common topological characteristics in a PPI network, including the degree, coreness, and closeness. Given such characteristics, PPI networks could be instrumental towards predicting the associations between proteins and diseases [12].

In this paper, we study the problem of disease protein prediction[1] exploiting network-based representations. Network-based approaches in this area generally fall into one of the three categories: linkage, module and diffusion-based methods [13]. Linkage methods are based on the assumption that the direct neighbors of a disease protein in a PPI network tend to be associated with the same disease. In particular, they focus on genomic linkage intervals [14, 15, 16]. If the protein products of the genes in a disease linkage interval interact with a known disease protein, then they become disease candidates. Module-based methods hypothesize that proteins within the same topological or functional module on a network are more likely to associate with the same disease [17, 10, 18]. In particular, there exist various approaches based on network clustering [19, 20], k-cores [21], graphlets [22], network motifs [23, 24] and frequent subgraph mining [25]. Finally, diffusion-based methods anchor on known disease proteins as seeds, which diffuse along PPI network through random walks [26, 27, 28, 29, 30].

However, PPI networks are often noisy and incomplete [31, 32]. Apart from capitalizing on these networks, most of the above methods do not consider the properties of proteins themselves, such as their Gene Ontology (GO) annotations like biological processes, molecular functions, cellular components, etc. For example, it is known that only the proteins which are localized at the same subcellular compartments can interact with each other [33, 34]. Thus, we propose to use keywords from the Universal Protein Resource (UniProt) database [35] to

[1] In this paper, we focus on disease gene prediction using the protein-protein interaction networks and thus disease genes and disease proteins are used interchangeably.

Table 1: A summary of keywords from the UniProt database.

Keyword Category	**Examples**
Biological Process	*Apoptosis, Cell cycle, cAMP biosynthesis*
Cellular component	*Golgi apparatus, Vacuole, Cytoplasm*
Coding sequence diversity	*Polymorphisms, RNA-editing, alternative splicing*
Domain	*SH2 domain, Kelch repeat, Transmembrane*
Ligand	*cAMP, S-adenosyl-L-methionine, cGMP*
Molecular function	*RNA-binding, Protein kinase inhibitor, Chromatin regulator*
Post-translational modification	*Phosphorylation, Ubiquitination, Acetylation*
Technical term	*Allosteric enzyme, Transposable element*

enrich the PPI network. The keywords cover various biological aspects of the proteins, as summarized in Table 1. A previous study [36] reveals the relationship between these keywords and intrinsic disorders: some keywords for cellular components, domains, technical terms, developmental processes, and coding sequence diversities indicate strong positive or negative correlations with long intrinsically disordered regions. Furthermore, it is known that intrinsically disordered proteins are associated with many diseases [37]. Additionally, several investigations [38, 39, 40] consider the role of post-translational modifications (PTM) in disease and functional complexes. Based on these findings, we integrate the *keywords* of Uniprot database directly into the PPI networks. With this integration we are able to capture the network characteristics between proteins and keywords as well. The concept of integrating additional biological knowledge into a PPI network has been studied in recent years [41, 42, 43, 44, 45, 46]. In particular, towards identifying disease-causing genes, Lage et al. have studied the computational integration of phenotype similarities to a PPI network in a pioneering work [47]. Another study [48] has generated a human disease network and disease gene network based on disease phoneme and genome associations. Moreover, Lee et al. [49] have improved the performance of genome-wide association studies in prioritizing candidate disease genes, through constructing a functional network for human genes based on various biological aspects such as mRNA coexpression, protein-protein interactions, and protein complex. Furthermore, a three-level network based on phenotype, protein complex and PPI network [50], as well as a heterogeneous network consisting of both interaction and ontology data, have been studied [51].

To the best of our knowledge, our work is the first attempt to integrate the PPI network with the *keywords* in Uniprot database and form a heterogeneous network for disease protein prediction. We call our heterogeneous network a *PPI-Keyword*

(PPIK) network, which contains network structures accounts for not only protein interactions with one another but also their functional and structural similarities. Based on the PPIK network, we propose to address the problem of disease protein classification, hinging on the notion of *metagraphs* [52]. Different from a traditional network motif or subgraph, a metagraph is a graph structure capturing a particular topology of both proteins and keywords on the PPIK network. In other words, each metagraph describes a particular heteronomous biological arrangement between one or more proteins and keywords. Each protein can be subsequently represented as a series of metagraphs that describe its interactions with other proteins and associations with keywords. The key intuition is that proteins with similar functional roles, such as their disease-causing property, tend to have similar metagraph representations, *i.e.*, they tend to interact with other proteins and associate with certain keywords in a similar arrangement on the PPIK network. Thus, we further build a classifier for disease proteins based on their metagraph representations. Finally, we conduct comprehensive experiments on three PPI databases, namely IntAct [53], STRING [54] and NCBI [55], and demonstrate the superior predictive power of our proposed metagraph-based prediction model.

2. Materials and Methods

In this section, we describe the proposed method. We start with some preliminaries in Section 2.1, including the problem statement as well as the motivations of our approach. Next, we introduce the proposed PPIK network and metagraph representations, the basis of our method in Section 2.2 and Section 2.3, respectively. Lastly, we present a general framework of our method in Section 2.4.

2.1. Preliminaries

Problem Statement. The problem of disease protein classification aims to identify human disease proteins in a given protein database. Let P be the protein space, and $C = \{\text{disease}, \text{non-disease}\}$ be the set of classes. Assume we have a training set P_{train} and test set P_{test} such that $P = P_{\text{train}} \cup P_{\text{test}}$ and $P_{\text{train}} \cap P_{\text{test}} = \emptyset$. The goal is to learn a classifier $\beta : P \rightarrow C$ based on P_{train}. Ultimately, for any protein in the test set $p \in P_{\text{test}}$, we can predict its class to be $\beta(p)$ with minimal prediction errors.

There have been numerous efforts in designing and learning different classifiers, with considerable success. All these classifiers generally assume a vector representation $\phi(p)$ for each protein $p \in P$. In this work, the main focus is

proposing a novel representation $\phi(\cdot)$ for proteins to consistently improve classifiers across the spectrum of classifiers, rather than developing a new classification model.

Motivations. First, while protein-protein interaction (PPI) networks have offered some insight into how proteins work with one another, most existing PPI networks are noisy and incomplete [31, 32]. Interestingly, an individual protein also exhibits a number of biological properties that may reveal how it works, ranging from molecular function and biological process to cellular component and protein domain. Previous work shows that disease proteins are likely to share common properties in Gene Ontology annotations [41, 56]. Furthermore, proteins with the same protein domain as a disease protein could be associated with the disease as well [57, 56]. Therefore, these properties can also be associated with different proteins together, to complement and enhance the existing PPI networks which only encode protein-protein interactions.

Second, disease proteins tend to have similar topological arrangements in a PPI network. This hypothesis has been validated to some extent by previous module-based methods such as clustering [19, 20], k-cores [21], graphlets [22], network motifs [23, 24] and frequent subgraph mining [25]. However, the challenge here is how we can leverage the topological arrangements of proteins in the context of not only the interactions between proteins but also their relevance based on other biological properties.

The two motivations inspire the proposed PPI-Keyword network in Section 2.2 and metagraph representation in Section 2.3, respectively.

2.2. PPI-Keyword network

To exploit the biological properties of individual proteins, we leverage the keywords associated with each protein from the Universal Protein Resource (UniProt) database [35]. These keywords describe the various biological mechanisms of the proteins, as summarized in Table 1. We enrich the PPI network with such biological keywords, to construct a PPI-Keyword (PPIK) network. Note that we only use the exact keywords, without considering the semantic similarity or overlap between keywords, since we find out that only 0.1% of the keyword pairs are similar.

Formally, a PPIK network is modeled by an undirected graph $G = (V, E, \ell)$, such that V is the set of nodes, E is the set of edges, and ℓ is the label function on V. In particular, each node can be either a protein or a keyword. They can be differentiated by the label function $\ell : V \to \{\text{protein}, \text{keyword}\}$. Furthermore,

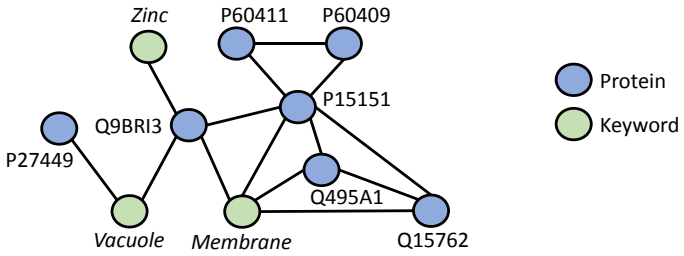


Figure 1: Part of the PPIK network based on UniProt and IntAct databases.

an edge can connect either two proteins, or a protein and a keyword. The former represents the mapped interactions between the two proteins, whereas the latter represents the association between the protein and keyword. Figure 1 shows a part of the constructed PPIK network based on the UniProt and IntAct [53] databases. Note that the color scheme of the nodes essentially serves as the label function.

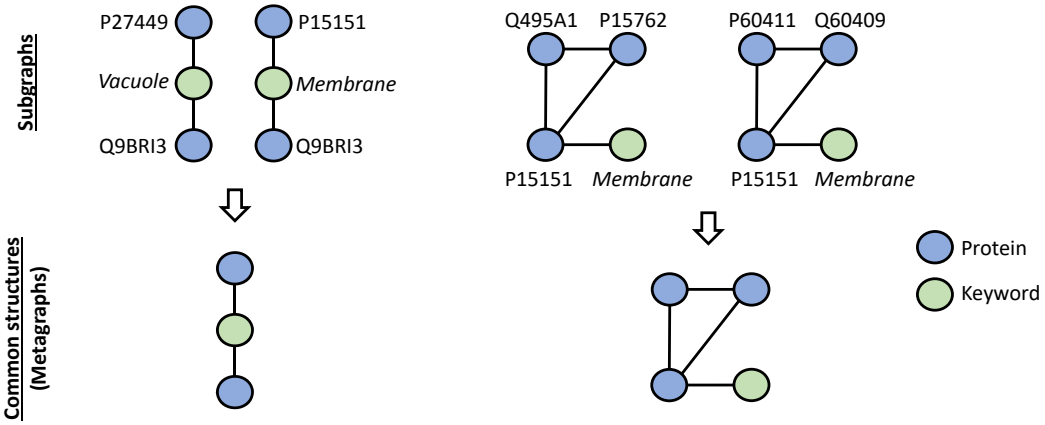
The PPIK network integrates keywords that describe biological mechanisms of proteins into the traditional PPI network. This integration complements the original noisy and incomplete network. On the one hand, protein-keyword associations could reinforce useful protein-protein interactions. On the other hand, proteins with no direct interactions can now become related through keywords.

2.3. Metagraph representations

On the PPIK network, proteins with similar roles (*e.g.*, their disease-causing functions) tend to have similar topological characteristics. To model topological similarities, previous module-based methods resort to structures such as network motifs [23, 24] and k-cores [21]. However, on a PPIK network, both proteins and keywords exist, and traditional structures do not differentiate between different labels of nodes. Fortunately, the recent emergence of metagraphs [52] has enabled the representation of common structures on a *heterogeneous* graph, where nodes with different labels connect with each other.

In the PPIK network as shown in Figure 1, we observe multiple subgraphs with a common structure, which are illustrated in Figure 2. More specifically, Figure 2(a) showcases two 3-node subgraphs of the PPIK network, both with a common structure "protein–keyword–protein". Likewise, Figure 2(b) illustrates two 4-node subgraphs with a common structure consisting of a triangle of three proteins and one keyword. We call such common structures metagraphs, and the corresponding subgraphs are their instances, i.e, metagraph instances.

Formally, a graph $S = (V_S, E_S, \ell)$ is a *subgraph* of graph $G = (V, E, \ell)$ iff $V_S \subseteq V$ and $E_S \subseteq E$. A graph $M = (V_M, E_M, \ell_M)$ is a *metagraph* for some label



(a) Example of 3-node metagraph (b) Example of 4-node metagraph

Figure 2: Example metagraphs: common structures of subgraphs on the PPIK network.

function ℓ_M, where each node is defined by its label and its value is immaterial. We say that S is an *instance* of M iff there exists a bijection ω between the nodes of S and M such that

- $\forall v \in V_s, \ell(v) = \ell_M(\omega(v))$, and
- $\forall v, u \in V_s, (v, u) \in E_S$ holds iff $(\omega(v), \omega(u)) \in E_M$ holds.

As a metagraph defines a specific topological arrangement of proteins and keywords, two proteins associated with the same metagraph tend to have similar functional roles. Therefore, we can use metagraphs to construct the vector representation of proteins. Let $\mathcal{M} \triangleq \{M_1, M_2, \ldots, M_{|\mathcal{M}|}\}$ denote the set of metagraphs on the PPIK network. Let $\mathcal{I}(M_i)$ be the set of instances of $M_i \in \mathcal{M}$. A protein p can be represented by a vector $\mathbf{m}_p$ of length $|\mathcal{M}|$, where the i-th element is the number of instances of M_i containing the protein p. That is,

$$\mathbf{m}_p[i] \triangleq |\{S \in \mathcal{I}(M_i) : p \in V_S\}|. \tag{1}$$

Furthermore, the same metagraph can have multiple subgraph instances of different "utilities". That is, a protein appearing in a subgraph together with a disease protein is more likely to be a disease protein, which implies that such subgraphs have a higher utility towards identifying disease proteins. As shown in Figure 3, some subgraphs contain disease proteins and some do not, based on biological knowledge from a disease database. To quantify such utilities, for each

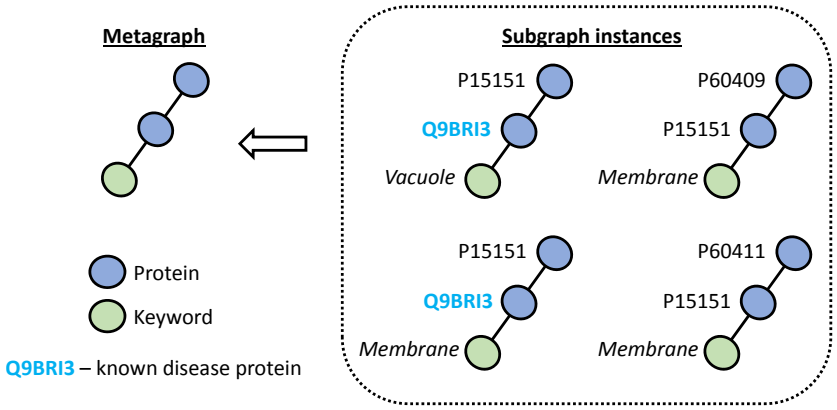


Figure 3: Example subgraph instances of a metagraph, where some contain disease proteins and some do not (Q9BRI3 is a known disease protein).

metagraph we compute the fraction of its subgraph instances containing any of the known disease proteins. The label function $\varphi : P \rightarrow \{\text{disease}, \text{non-disease}\}$ differentiates known disease proteins from other proteins. Formally, let $\mathbf{d}_p$ be a vector of length $|\mathcal{M}|$, where the i-th element is defined as follows:

$$\mathbf{d}_p[i] \triangleq \frac{|\{S \in \mathcal{I}(M_i) : p \in V_S \wedge (\exists v \in V_s : v \neq p \wedge \varphi(v) = \text{disease})\}|}{\mathbf{m}_p[i]}. \quad (2)$$

$\mathbf{m}_p$ and $\mathbf{d}_p$ are $|\mathcal{M}|$-dimensional representations of protein p based on *metagraphs*. In addition, keywords describing each protein naturally become part of its vector representation. Given a set of keywords $\mathcal{K}$, let $\mathbf{k}_p$ be a vector of length $|\mathcal{K}|$, where the i-th element is 1 iff the i-th keyword is associated with protein p. Thus, $\phi(p)$, the *overall* vector representation of protein p is a vector with $(2|\mathcal{M}| + \mathcal{K})$ dimensions, which is the concatenation of the above representations as shown in Equation (3).

$$\phi(p) \triangleq [\mathbf{m}_p, \mathbf{d}_p, \mathbf{k}_p]. \quad (3)$$

2.4. General framework

Based on the proposed PPIK network and metagraph representations, we describe an overall framework to learn a classifier for disease protein prediction. In particular, the framework consists of three main steps, as summarized in Figure 4.

First, from the PPIK network, we mine the collection of metagraphs $\mathcal{M}$. This is an active research area and many off-the-shelf solutions exist. Therefore, we

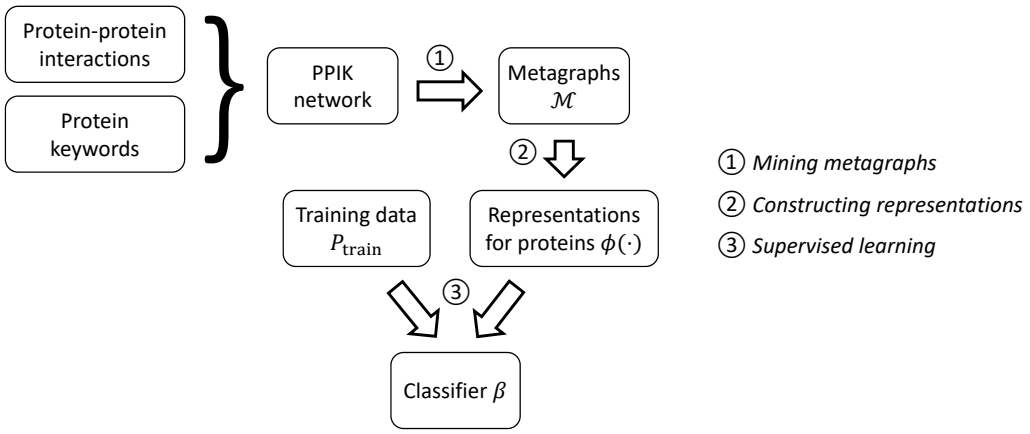


Figure 4: General framework of the proposed method.

apply an existing state-of-the-art approach GRAMI [58] for this step. In particular, we only consider metagraphs up to 5 nodes, which is a good balance between efficiency and accuracy.

Second, we derive the *metagraph* representations for proteins based on the mined metagraphs $\mathcal{M}$. We employ the SymISO algorithm [52] to compute the set of instances $\mathcal{I}(M_i)$ for each metagraph $M_i \in \mathcal{M}$ and then construct the metagraph representations $\mathbf{m}_p$ and $\mathbf{d}_p$ as defined in Eq. 1–2.

Third, based on the protein representations $\phi(\cdot)$ and training data P_{train}, we build a classifier β through supervised learning. Note that the main focus of this work is to propose the metagraph representations based on the PPIK network, which aims to improve disease prediction across various supervised learning techniques including random forest, SVM and Generalized Linear Models, as we will demonstrate in the experiments.

3. Results and Discussion

In this section, we empirically evaluate the effect of metagraph representations in the context of disease protein prediction. Results show that the proposed representations can significantly improve prediction across various classifiers and substantially outperforms random walk baselines RWR [29] and PRINCE [59]. The reason of choosing these diffusion/propagation based methods is their dominant power over the clustering and neighborhood methods [60, 61, 62].

3.1. Data and setup

In this paper, to demonstrate the effects of our proposed novel protein representation, we work on three different human PPI databases, namely IntAct [53], NCBI [55] and STRING [54]. We also exploit protein keywords from the UniProt database [35] as illustrated in Table 1, to construct a PPIK network for each PPI database. Disease labels for proteins are obtained from the UniProt and OMIM databases [63]. In particular, we first obtain disease genes from OMIM, and further map these genes to their product proteins based on UniProt. We conduct disease protein prediction under two different scenarios as follows. First, whether a protein is associated with *all disease*, i.e, all phenotypes in OMIM; second, whether a protein is specifically associated with *breast cancer*, i.e, phenotype breast cancer. Table 2 summarizes the three PPIK networks. As we can see from Table 2, the three PPIK networks are very different in terms of number of proteins, number of PPI edges, as well as number of PPIK edges.

Table 2: Summary of the three PPIK networks.

	Proteins	**Disease proteins**		**Keywords**	**PPI edges**	**PPIK edges**
		All	**Breast cancer**			
IntAct	13 063	2 947	29	554	97 652	246 092
NCBI	15 951	3 476	31	567	227 004	405 632
STRING	17 668	3 539	29	567	3 912 853	4 107 335

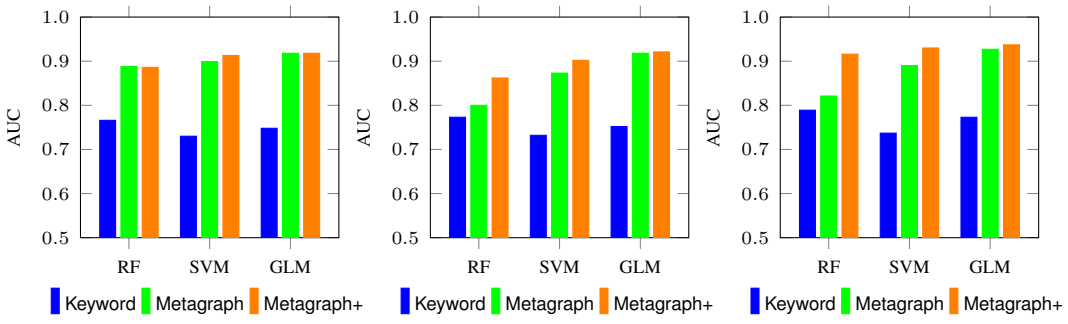
We split each dataset into training and testing sets, containing 80% and 20% proteins respectively. The split is repeated 5 different times. All results reported are averaged over the 5 splits.

To evaluate the effectiveness of the proposed method, we employ the standard metric of Area Under the ROC Curve (AUC), which is a robust measure of the classifiers' predictive power even with unbalanced classes (*e.g.*, breast cancer).

3.2. Benefits of metagraph-based representations

As discussed in Section 2, given a protein representation $\phi(\cdot)$, different supervised learning models can be applied to derive a disease protein classifier. In particular, we consider three progressively richer representations as follows.

- Keyword. We only consider protein keywords from UniProt database (see Table 1), *i.e.*, $\phi(p) = \mathbf{k}_p$. These keywords describe various biological properties of proteins, and thus encompass reasonable predictive power for disease proteins.



(a) IntAct (b) NCBI (c) STRING

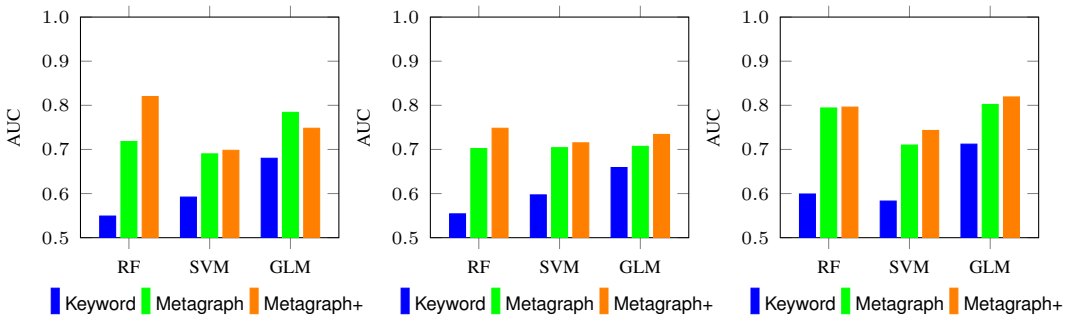
Figure 5: Performance of Metagraph and Metagraph+ compared to Keywords for *all disease*.

- Metagraph. We enhance the keyword-based representation with metagraph statistics (see Eq. 1), *i.e.*, $\phi(p) = [\mathbf{k}_p, \mathbf{m}_p]$. The vector $\mathbf{m}_p$ captures the topological arrangement on the PPIK network for interactions between both proteins and keywords.

- Metagraph+. We further incorporate metagraph representations based on the utilities of their subgraph instances (see Eq. 2), *i.e.*, $\phi(p) = [\mathbf{k}_p, \mathbf{m}_p, \mathbf{d}_p]$. In particular, $\mathbf{d}_p$ differentiates metagraphs based on the disease class of the proteins in their subgraph instances.

The above representations are meant to work across different models of supervised learning. In our experiments, we adopted 3 well-known classification models, namely, Random Forest (**RF**), Support Vector Machine (**SVM**) and Generalized Linear Model (**GLM**). For RF, we used `randomforest` package in R and tuned the mtry parameter between 1 and the cardinality of protein representation with `tuneRF` function based on OOB error. For SVM, we used `e1071` package in R, and tuned the gamma parameter over $\{10^{-5}, 10^{-4}, 0.001, 0.01, 0.1\}$ and the cost parameter over $\{0.1, 1, 10\}$ with grid-based `tune.svm` function (based on classification error). For GLM, we used `stats` package in R and adopted default Gaussian distribution. Furthermore, in the case of *breast cancer*, the classes are highly unbalanced. Therefore, we used the `ROSE` package in R to oversample the *breast cancer* class with probability 0.2.

Figure 5 shows the AUC performance of the three representations in each of the classification method, on each of the three datasets, for *all diseases*. Two main observations can be made from the results. First, the metagraph representation (Metagraph) can significantly and consistently outperform the keyword-only



(a) IntAct (b) NCBI (c) STRING

Figure 6: Performance of Metagraph and Metagraph+ compared to Keywords for *breast cancer*.

representation (Keyword) across all circumstances. Averaging over all classifiers and datasets, Metagraph improves AUC over Keyword by 12.6%. The results imply that interactions/associations on the PPIK network are powerful towards disease prediction; in particular, proteins with similar functional roles tend to appear in similar topological arrangements. On the contrary, it is inadequate to only consider keywords for individual proteins. Second, the utility-based metagraph representation Metagraph+ can further enhance the performance. Overall, Metagraph+ can achieve an average AUC of 90.9, as compared to 88.2 for Metagraph and 75.6 for Keyword. Thus, metagraphs can effectively incorporate different utilities based on disease proteins in the subgraph instances.

Next, we zoom into the results for *breast cancer* only, as shown in Figure 6. The performance differences of Keyword, Metagraph and Metagraph+ are similar to those for all diseases. Averaging across all classifiers and datasets, Metagraph+ attains an AUC of 75.8, beating Metagraph and Keyword by 2.3% and 14.4%, respectively. The results reaffirm that disease proteins tend to appear in the same neighborhood of the PPIK network, and our metagraph representations carry strong predictive power.

3.3. Comparison to baselines

Having demonstrated the benefits of metagraph representations, we now compare our proposed work to other baselines for protein disease prediction, as follows.

- RWR or Random walk with Restart [29]. On the PPI network, consider a particle initially at one of the known disease proteins, *i.e.*, the initial position of the particle has a uniform distribution over the disease proteins in the

training data. Next, in each step, the particle makes a move on the network: either moving to a randomly selected neighbor with $1 - \alpha$ probability, or jumping to one of the disease proteins in the training data with α probability. Note that the "jumping" effectively returns the particle to the initial condition, and thus "restarts" the random walk. The process is repeated until it converges to a stationary distribution over all the proteins. In the end, candidate proteins in the test data are ranked according to the stationary distribution. We chose `RANKS` package in R as the implementation. The α parameter is tuned over $\{0.1, 0.2, \ldots, 0.9\}$ based on AUC performances.

- RWRK. The same method as RWR, except that the random walk with restart is performed on the PPIK network.

- PRINCE [59]. In our case, i.e, on an unweighted PPI network, this method basically performs random walk with restart with one major difference than RWR that is prior probabilities. In RWR the initial probability vector consist of equal probabilities apportioned between known disease associated proteins. On the other hand, in PRINCE prior probabilities are assigned to each disease associated protein based on a logistic function: $L(x) = \frac{1}{1 + e^{(cx+d)}}$ and $x = S(q, p)$ where $S(q, p)$ is the similarity score between query disease q and the associated disease p with the protein. If the protein is associated with more than one disease then, p is chosen to be the most similar one to q. We use the recommended values as in paper [59] and set c to -15 and d to $log(9999)$. The similarity scores between phenotypes are derived from the study [64]. The α restart probability parameter is tuned over $\{0.1, 0.2, \ldots, 0.9\}$ based on AUC performances as in RWR.

- PRINCEK. The same method as PRINCE, except that the algorithm is performed on the PPIK network.

- Subgraph+. We also compare proposed work to subgraph-based approaches. Unlike metagraphs, traditional subgraphs or network motifs do not differentiate heterogeneous types of nodes. In this baseline, we consider subgraphs with only protein nodes, and their statistics and utilities are formulated as protein representations similar to the case of metagraphs (see Eq. 1–2), which are then concatenated with keyword representations for individual proteins. We call this method Subgraph+, analogous to Metagraph+. We leveraged Subgraph+ by different classifiers with the same tunning routine as Metagraph and Metagraph+.

Table 3: Performance of Metagraph+ compared to random walk and subgraph baselines.

	All Disease			**Breast Cancer**		
	IntAct	**NCBI**	**STRING**	**IntAct**	**NCBI**	**STRING**
RWR	0.551	0.567	0.622	0.578	0.665	0.605
RWRK	0.590	0.587	0.629	0.587	0.664	0.612
PRINCE	-	-	-	0.506	0.716	0.632
PRINCEK	-	-	-	0.634	0.717	0.596
Classifier: RF						
Subgraph+	0.745	0.756	0.826	0.611	0.696	0.713
Metagraph+	**0.886**	**0.862**	**0.916**	**0.820**	**0.748**	**0.796**
Classifier: SVM						
Subgraph+	0.739	0.741	0.808	0.687	0.682	0.597
Metagraph+	**0.913**	**0.902**	**0.930**	**0.698**	0.715	**0.743**
Classifier: GLM						
Subgraph+	0.751	0.758	0.818	0.733	0.715	0.796
Metagraph+	**0.918**	**0.921**	**0.937**	**0.748**	**0.734**	**0.819**

Table 3 compares the AUC of the baseline methods with Metagraph+, for *all diseases* and *breast cancer*. The foremost observation is that with classifiers GLM and RF, Metagraph+ is significantly better than all baselines. On average considering all the classifiers, for *all diseases*, it outperforms RWR, RWRK, and Subgraph+ by 32.9%, 30.7% and 13.8% respectively; for *breast cancer* it outperforms RWR, RWRK, PRINCE, PRINCEK and Subgraph+ by 14.2%, 13.7%, 14%, 11% and 6.6% respectively. We attribute the better performance to the more predictive representations enabled by metagraphs. Second, the results also show that PPIK network is more effective than PPI network, as evident from the comparison between RWR and RWRK methods. More specifically, random walks on the PPIK network (RWRK) produce more accurate predictions than those on the PPI network (RWR) under most circumstances. In other words, protein keywords can indeed complement and enrich the PPI network.

Table 4 further examines the running times of the baselines RWR, RWRK, PRINCE, PRINCEK, Subgraph+ and the proposed approach(es). It can be concluded that for RWR and PRINCE, *keywords* integrated into the PPI network induced an increase in the running times. Moreover, for Metagraph+ we observe unfavorable effect of additional vector $\mathbf{d}_p$ of length $|\mathcal{M}|$ (metagraph set size) compare to Metagraph.

Table 4: Running time (train and test in minutes) comparisons. For Subgraph+, Metagraph, and Metagraph+ the classifier with the best AUC performance, GLM, is chosen.

	All Disease			**Breast Cancer**		
	IntAct	**NCBI**	**STRING**	**IntAct**	**NCBI**	**STRING**
RWR	1.0	1.3	2.0	1.4	1.5	1.0
RWRK	1.4	2.3	3.9	1.4	2.2	2.9
PRINCE	-	-	-	0.7	1.0	1.1
PRINCEK	-	-	-	0.9	1.4	1.6
Subgraph+	0.4	1.0	0.8	0.6	0.6	0.8
Metagraph	0.2	0.2	0.2	0.4	0.4	0.2
Metagraph+	0.8	1.0	1.6	1.4	2.0	2.6

3.4. Further analysis of the predicted disease proteins

We further study the disease proteins predicted by our proposed methods. Since GLM generally has the best performance among the three classifiers, we only focus on the results from this classifier. A protein is classified as a disease protein if its prediction score is higher than 0.5. Moreover, for each proposed method, we ran disease classification on all three datasets (IntAct, NCBI and STRING). Then, we combine the predictions for all three datasets to obtain a predicted disease gene set for each method.

To enable the analysis, DisGeNET database [65] is used to search the PubMed Ids of the up-to-date publications[2] reporting the gene-disease associations. In particular, we transform our predicted disease proteins to their producer gene Id's based on UniProt. Figure 7 illustrates the average number of publications per prediction that support the disease gene predictions of our proposed methods. The results are consistent with the AUC performance reported earlier, where methods with higher average publications attain higher AUC scores. Table 5 further lists the 10 genes with the most number of PubMed publications for our best proposed method, Metagraph+. Table 6 zooms into each of the 10 genes, and illustrates the top diseases associated with each gene based on DisGeNET scoring (which combines both curated content and literature).

We also evaluate novel proteins predicted by Metagraph+, based on recent publications from years 2014–2016, where each prediction has fewer than 20 publications. There are 27 such novel genes predicted (mapped from the predicted proteins), as listed in Table 7. For example, PDE9A (GeneId: 5152) has been

[2]Up to the year 2016, when we conducted this analysis.

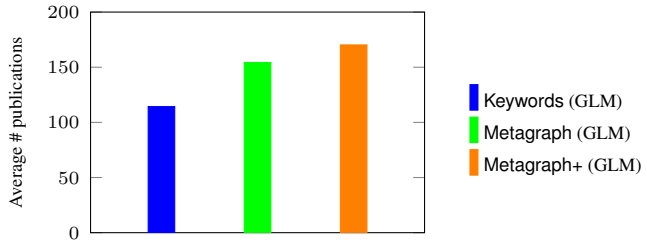


Figure 7: Average number of PubMed publications per prediction based on DisGeNET.

Table 5: Top 10 genes predicted by Metagraph+ (GLM) with the most number of publications.

GeneId	Symbol	Full Name
3565	IL4	interleukin 4
3552	IL1A	interleukin 1, alpha
4513	COX2	cytochrome c oxidase subunit II
3133	HLA-E	major histocompatibility complex, class I, E
6696	SPP1	secreted phosphoprotein 1
2272	FHIT	fragile histidine triad
7298	TYMS	thymidylate synthetase
1813	DRD2	dopamine receptor D2
2100	ESR2	estrogen receptor 2 (ER beta)
2078	ERG	v-ets avian erythroblastosis virus E26 oncogene homolog

reported to be associated with heart failure [66], and ZNF366 (GeneId: 167465) has been reported to be associated with estrogen metabolism and progression of breast cancer, and a new candidate for endometriosis as well [67]. Furthermore, a study [68] presents supportive evident for KCND2 (GeneId: 3751) being a causal gene for epilepsy, whereas another study [69] proposes KCND2 as a novel cause of J-wave syndrome associated with sudden cardiac arrest.

Table 6: Top 5 associated diseases reported by DisGeNET for the 10 genes reported in Table 5.

Disease UMLS and Name	**#PM**	**Disease UMLS and Name**	**#PM**
GeneId: 3565		GeneId: 3552	
C0004096 Asthma	199	C0018843 Heat Stroke	2
C0011615 Dermatitis, Atopic	65	C0035126 Reperfusion Injury	2
C0034069 Pulmonary Fibrosis	6	C0021368 Inflammation	17
C0035455 Rhinitis	8	C0011633 Dermatomyositis	3
C0993582 Arthritis, Experimental	4	C0037274 Dermatologic disorders	8
GeneId: 4513		GeneId: 3133	
C0268237 Cytochrome-c Oxidase Deficiency	5	C0030491 Parapsoriasis	1
C0027819 Neuroblastoma	5	C0151744 Myocardial Ischemia	2
C0013080 Down Syndrome	1	C0011854 Diabetes Mellitus, Insulin-Dependent	107
C0011853 Diabetes Mellitus, Experimental	1	C0004364 Autoimmune Diseases	105
C0151786 Muscle Weakness	1	C0019693 HIV Infections	86
GeneId: 6696		GeneId: 2272	
C0022650 Kidney Calculi	9	C0024121 Lung Neoplasms	25
C0017638 Glioma	15	C0038356 Stomach Neoplasms	5
C0006663 Calcinosis	3	C0033578 Prostatic Neoplasms	4
C0027627 Neoplasm Metastasis	147	C0007131 Non-Small Cell Lung Carcinoma	29
C1458155 Mammary Neoplasms	18	C0025500 Mesothelioma	2
GeneId: 7298		GeneId: 1813	
C0009404 Colorectal Neoplasms	47	C0036341 Schizophrenia	162
C0038356 Stomach Neoplasms	17	C1834570 Myoclonic dystonia	6
C0009375 Colonic Neoplasms	19	C0030567 Parkinson Disease	22
C1458155 Mammary Neoplasms	10	C0236736 Cocaine-Related Disorders	13
C0034885 Rectal Neoplasms	9	C0001973 Alcoholic Intoxication, Chronic	163
GeneId: 2100		GeneId: 2078	
C1458155 Mammary Neoplasms	48	C0033578 Prostatic Neoplasms	37
C0282612 Prostatic Intraepithelial Neoplasias	2	C0023467 Leukemia, Myelocytic, Acute	34
C0024668 Mammary Neoplasms, Experimental	2	C0023418 leukemia	18
C0014175 Endometriosis	23	C0553580 Ewings sarcoma	22
C0033578 Prostatic Neoplasms	18	C0013080 Down Syndrome	10

Table 7: Novel disease genes (mapped from predicted proteins) discovered by Metagraph+ based on recent PubMed publications from years 2014–2016.

GeneId	Symbol	PubMedId	Diseases
319	APOF	25726912	Liver neoplasms; Liver carcinoma
506	ATP5B	25666834	Non-alcoholic Fatty Liver Disease; Acute kidney injury
988	CDC5L	26089329	Ischemic Cerebrovascular Accident; Ischemic stroke
1635	DCTD	25735499	Neutropenia; Leukopenia
2686	GGT7	25884624	Glioblastoma; Glioma; Carcinogenesis
3751	KCND2	24501278	Epilepsy; Autistic Disorder; Seizures
		25214526	Cardiac Arrest
		25878292	Alzheimer's Disease
5152	PDE9A	25799991	Heart failure; Heart Diseases; Congestive heart failure
5634	PRPS2	25149475	Lupus Erythematosus, Systemic
		26004865	Congenital absence of germinal epithelium of testes
6723	SRM	25889691	Prostate carcinoma; Malignant neoplasm of prostate
6942	TCF20	25228304	Autism Spectrum Disorders; Atrial Septal Defects; Moderate mental retardation
8974	P4HA2	25741866	Severe myopia
		26001784	Disorder of skeletal system
9620	CELSR1	25117632	Ischemic stroke
10584	COLEC10	25495265	Chronic Lymphocytic Leukemia
		25786252	Chronic Lymphocytic Leukemia
10940	POP1	26275995	Inflammatory disorder
11097	NUPL2	25584925	Chronic Obstructive Airway Disease; Chronic Obstructive Airway Disease
22938	SNW1	26103569	Skin carcinoma
23513	SCRIB	24802235	Neoplasm Metastasis
23710	GABARAPL1	24879149	Malignant neoplasm of breast; Breast Carcinoma
26270	FBXO6	25811541	Stevens-Johnson Syndrome
55576	STAB2	25989359	Ankylosing spondylitis
84870	RSPO3	24430505	Osteoporosis
115426	UHRF2	25664994	Hepatitis B
144568	A2ML1	26121085	Otitis Media
162515	SLC16A11	25839936	Diabetes Mellitus, Non-Insulin-Dependent; Diabetes; Diabetes Mellitus
		25973943	Gestational Diabetes; Diabetes Mellitus, Non-Insulin-Dependent
167465	ZNF366	25722978	Breast Carcinoma; Malignant neoplasm of breast; Endometriosis; Endometrioma
285671	RNF180	24833402	Stomach Carcinoma; Malignant neoplasm of stomach
340419	RSPO2	25769727	Pancreatic carcinoma; Malignant neoplasm of pancreas

4. Conclusion

Disease protein prediction is crucial to the diagnosis and treatment of many diseases. In this study, we integrated protein-protein interaction and biological keywords of proteins, to construct a novel PPIK network. Based on the PPIK network, we further proposed metagraph representations for proteins. Such novel representations can improve the classification of disease proteins consistently across different classifiers, outperforming them by 15.3% in AUC on average. Our method also beats random walk and subgraph baselines by 13.8–32.9%. Finally, our literature search based on PubMed revealed that the proposed method can indeed better predict disease proteins that mapped with newly discovered biological knowledge.

5. Acknowledgements

The authors thank Dr. Wenqing Lin, for his initial valuable discussions.

References

[1] M. R. Nelson, H. Tipney, J. L. Painter, J. Shen, P. Nicoletti, Y. Shen, A. Floratos, P. C. Sham, M. J. Li, J. Wang, L. R. Cardon, J. C. Whittaker, P. Sanseau, The support of human genetic evidence for approved drug indications, Nat Genet 47 (2015) 856–860.

[2] W. A. Flavahan, Y. Drier, B. B. Liau, S. M. Gillespie, A. S. Venteicher, A. O. Stemmer-Rachamimov, M. L. Suvá, B. E. Bernstein, Insulator dysfunction and oncogene activation in idh mutant gliomas, Nature 529 (2016) 110–114.

[3] A. Sekar, A. R. Bialas, H. de Rivera, A. Davis, T. R. Hammond, N. Kamitaki, K. Tooley, J. Presumey, M. Baum, V. Van Doren, G. Genovese, S. A. Rose, R. E. Handsaker, S. W. G. o. t. P. G. Consortium, M. J. Daly, M. C. Carroll, B. Stevens, S. A. McCarroll, Schizophrenia risk from complex variation of complement component 4, Nature 530 (2016) 177–183.

[4] P. Singh, J. C. Schimenti, The genetics of human infertility by functional interrogation of snps in mice, Proc Natl Acad Sci U S A 112 (2015) 10431–10436.

[5] P. Yang, X. Li, H.-N. Chua, C.-K. Kwoh, S.-K. Ng, Ensemble positive unlabeled learning for disease gene identification, PLOS ONE 9 (2014) 1–11.

[6] P. Yang, X.-L. Li, J.-P. Mei, C.-K. Kwoh, S.-K. Ng, Positive-unlabeled learning for disease gene identification, Bioinformatics 28 (2012) 2640.

[7] M. Li, Y. Lu, J. Wang, F.-X. Wu, Y. Pan, A topology potential-based method for identifying essential proteins from ppi networks, IEEE/ACM Transactions on Computational Biology and Bioinformatics 12 (2015) 372–383.

[8] T. Gui, X. Dong, R. Li, Y. Li, Z. Wang, Identification of hepatocellular carcinoma-related genes with a machine learning and network analysis, Journal of Computational Biology 22 (2015) 63–71.

[9] L. Fu, S. Zhang, L. Zhang, X. Tong, J. Zhang, Y. Zhang, L. Ouyang, B. Liu, J. Huang, Systems biology network-based discovery of a small molecule activator bl-ad008 targeting ampk/zipk and inducing apoptosis in cervical cancer, Oncotarget 6 (2015) 8071–8088.

[10] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, T. Ideker, Network-based classification of breast cancer metastasis, Molecular Systems Biology 3 (2007) 140–n/a.

[11] T. Ideker, R. Sharan, Protein networks in disease, Genome Research 18 (2008) 644–652.

[12] J. Xu, Y. Li, Discovering disease-genes by topological features in human protein–protein interaction network, Bioinformatics 22 (2006) 2800–2805.

[13] A. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: a network-based approach to human disease, Nat Rev Genet 12 (2011).

[14] S. Kathiresan, M. I. G. Consortium, B. Voight, Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants, Nat Genet 41 (2009).

[15] I. Iossifov, T. Zheng, M. Baron, T. C. Gilliam, A. Rzhetsky, Genetic-linkage mapping of complex hereditary disorders to a whole-genome molecular-interaction network, Genome Research 18 (2008) 1150–1162.

[16] M. Krauthammer, C. A. Kaufmann, T. C. Gilliam, A. Rzhetsky, Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in alzheimer's disease, Proceedings of the National Academy of Sciences of the United States of America 101 (2004) 15148–15153.

[17] S. Suthram, J. T. Dudley, A. P. Chiang, R. Chen, T. J. Hastie, A. J. Butte, Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets, PLOS Computational Biology 6 (2010) 1–10.

[18] G. Wu, L. Stein, A network module-based method for identifying cancer prognostic signatures, Genome Biology 13 (2012) R112.

[19] A. King, N. Pržulj, I. Jurisica, Protein complex prediction via cost-based clustering, Bioinformatics 20 (2004) 3013–3020.

[20] Y.-K. Shih, S. Parthasarathy, Identifying functional modules in interaction networks through overlapping markov clustering, Bioinformatics 28 (2012) i473–i479.

[21] Y. Zhang, Z. Li, M. Yang, D. Wang, L. Yu, C. Guo, X. Guo, N. Lin, Identification of grb2 and gab1 coexpression as an unfavorable prognostic factor for hepatocellular carcinoma by a combination of expression profile and network analysis, PLoS ONE 8 (2013).

[22] T. Milenkovi, N. Pržulj, Uncovering biological network function via graphlet degree signatures, Cancer Informatics 6 (2008) 257–273.

[23] Y. R. Cho, Y. Xin, G. Speegle, P-finder: Reconstruction of signaling networks from protein-protein interactions and go annotations, IEEE/ACM Transactions on Computational Biology and Bioinformatics 12 (2015) 309–321.

[24] N. T. L. Tran, S. Mohan, Z. Xu, C.-H. Huang, Current innovations and future challenges of network motif detection, Briefings in Bioinformatics 16 (2015) 497.

[25] Y.-R. Cho, A. Zhang, Predicting protein function by frequent functional association pattern mining in protein interaction networks, Trans. Info. Tech. Biomed. 14 (2010) 30–36.

[26] Y. Li, J. C. Patra, Genome-wide inferring genephenotype relationship by walking on the heterogeneous network, Bioinformatics 26 (2010) 1219–1224.

[27] G. U. Ganegoda, J. Wang, F. X. Wu, M. Li, Prioritization of candidate genes based on disease similarity and protein's proximity in ppi networks, in: 2013 IEEE International Conference on Bioinformatics and Biomedicine, pp. 103–108.

[28] X. Wu, R. Jiang, M. Q. Zhang, S. Li, Network-based global inference of human disease genes, Mol Syst Biol 4 (2008) 189–189.

[29] S. Köhler, S. Bauer, D. Horn, P. N. Robinson, Walking the interactome for prioritization of candidate disease genes, The American Journal of Human Genetics 82 (2008) 949 – 958.

[30] D.-H. Le, Network-based ranking methods for prediction of novel disease associated micrornas, Computational Biology and Chemistry 58 (2015) 139–148.

[31] L. Zhu, S.-P. Deng, D.-S. Huang, A two-stage geometric method for pruning unreliable links in protein-protein networks, IEEE Transactions on Nanobioscience 14 (2015) 528–534.

[32] P. Marcatili, A. Tramontano, Network cleansing: Reliable interaction networks, IGI Global, pp. 80–97.

[33] X. Tang, X. Hu, X. Yang, Y. Fan, Y. Li, W. Hu, Y. Liao, M. c. Zheng, W. Peng, L. Gao, Predicting diabetes mellitus genes via protein-protein interaction and protein subcellular localization information, BMC Genomics 17 (2016) 433.

[34] J. X. Binder, S. Pletscher-Frankild, K. Tsafou, C. Stolte, S. I. O'Donoghue, R. Schneider, L. J. Jensen, Compartments: unification and visualization of protein subcellular localization evidence, Database (Oxford) 2014 (2014) bau012.

[35] U. Consortium, et al., Uniprot: a hub for protein information, Nucleic acids research (2014) gku989.

[36] S. Vucetic, H. Xie, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic, V. N. Uversky, Functional anthology of intrinsic disorder. 2. cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions, Journal of Proteome Research 6 (2007) 1899–1916.

[37] V. N. Uversky, C. J. Oldfield, A. K. Dunker, Intrinsically disordered proteins in human diseases: Introducing the d2 concept, Annual Review of Biophysics 37 (2008) 215–246.

[38] T. M. Karve, A. K. Cheema, Small changes huge impact: The role of protein posttranslational modifications in cellular homeostasis and disease, J Amino Acids 2011 (2011) 207691.

[39] M. Mann, O. N. Jensen, Proteomic analysis of post-translational modifications, Nat Biotech 21 (2003) 255–261.

[40] J. Woodsmith, U. Stelzl, A. Vinayagam, Bioinformatics Analysis of PTM-Modified Protein Interaction Networks and Complexes, Springer New York, New York, NY, pp. 321–332.

[41] K. Sun, J. P. Gonçalves, C. Larminie, N. Pržulj, Predicting disease associations via biological network analysis, BMC Bioinformatics 15 (2014) 304.

[42] W. Liu, A. Wu, M. Pellegrini, X. Wang, Integrative analysis of human protein, function and disease networks, Scientific Reports 5 (2015) 14344 EP –. Article.

[43] U. M. Singh-Blom, N. Natarajan, A. Tewari, J. O. Woods, I. S. Dhillon, E. M. Marcotte, Prediction and validation of gene-disease associations using methods inspired by social network analyses, PLOS ONE 8 (2013) 1–17.

[44] W. Peng, J. Wang, J. Cai, L. Chen, M. Li, F.-X. Wu, Improving protein function prediction using domain and protein complexes in ppi networks, BMC Syst Biol 8 (2014) 35–35. 1752-0509-8-35[PII].

[45] Z. H. Yang, F. Y. Yu, H. F. Lin, J. Wang, Integrating ppi datasets with the ppi data from biomedical literature for protein complex detection, BMC Med Genomics 7 (2014) S3–S3. 1755-8794-7-S2-S3[PII].

[46] A. Ruepp, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, M. Stransky, B. Waegele, T. Schmidt, O. N. Doudieu, V. Stmpflen, H. W. Mewes, Corum: the comprehensive resource of mammalian protein complexes, Nucleic Acids Research 36 (2008) D646.

[47] K. Lage, E. O. Karlberg, Z. M. Storling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tumer, F. Pociot, N. Tommerup, Y. Moreau, S. Brunak, A human phenome-interactome network of protein complexes implicated in genetic disorders, Nat Biotech 25 (2007) 309–316.

[48] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, A. Barabási, The human disease network, Proceedings of the National Academy of Sciences 104 (2007) 8685–8690.

[49] 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 Res 21 (2011) 1109–1121. 21536720[pmid].

[50] P. Yang, X. Li, M. Wu, C.-K. Kwoh, S.-K. Ng, Inferring gene-phenotype associations via global protein complex network propagation, PLOS ONE 6 (2011) 1–11.

[51] J. Peng, K. Bai, X. Shang, G. Wang, H. Xue, S. Jin, L. Cheng, Y. Wang, J. Chen, Predicting disease-related genes using integrated biomedical networks, BMC Genomics 18 (2017) 1043.

[52] Y. Fang, W. Lin, V. W. Zheng, M. Wu, K. C. Chang, X. Li, Semantic proximity search on graphs with metagraph-based learning, in: 32nd IEEE International Conference on Data Engineering, ICDE 2016, Helsinki, Finland, May 16-20, 2016, pp. 277–288.

[53] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. del Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R. C. Lovering, B. Meldal, A. N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni, H. Hermjakob, The mintact project intact as a common curation platform for 11 molecular interaction databases, Nucleic acids research 42 (2014) D358–63.

[54] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, C. von Mering, String v10: protein-protein interaction networks, integrated over the tree of life, Nucleic Acids Research 43 (2015) D447.

[55] D. Maglott, J. Ostell, K. D. Pruitt, T. Tatusova, Entrez gene: gene-centered information at ncbi, Nucleic Acids Research 35 (2007) D26.

[56] R. Jiang, M. Wu, L. Li, Pinpointing disease genes through phenomic and genomic data fusion, BMC Genomics 16 (2015) S3.

[57] J.-S. Chen, W.-S. Hung, H.-H. Chan, S.-J. Tsai, H. S. Sun, In silico identification of oncogenic potential of fyn-related kinase in hepatocellular carcinoma, Bioinformatics 29 (2013) 420.

[58] M. Elseidy, E. Abdelhamid, S. Skiadopoulos, P. Kalnis, Grami: Frequent subgraph and pattern mining in a single large graph, Proc. VLDB Endow. 7 (2014) 517–528.

[59] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, R. Sharan, Associating genes and protein complexes with disease via network propagation, PLOS Computational Biology 6 (2010) 1–9.

[60] S. Navlakha, C. Kingsford, The power of protein interaction networks for associating genes with diseases, Bioinformatics 26 (2010) 1057.

[61] J. Zhu, Y. Qin, T. Liu, J. Wang, X. Zheng, Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles, BMC Bioinformatics 14 (2013) S5.

[62] J. E. Shim, S. Hwang, I. Lee, Pathway-dependent effectiveness of network algorithms for gene prioritization, PLOS ONE 10 (2015) 1–10.

[63] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, V. A. McKusick, Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders, Nucleic Acids Research 30 (2002) 52.

[64] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, J. A. M. Leunissen, A text-mining analysis of the human phenome, Eur J Hum Genet 14 (2006) 535–542.

[65] J. Piero, l. Bravo, N. Queralt-Rosinach, A. Gutirrez-Sacristn, J. Deu-Pons, E. Centeno, J. Garca-Garca, F. Sanz, L. I. Furlong, Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants, Nucleic Acids Research 45 (2017) D833.

[66] D. I. Lee, G. Zhu, T. Sasaki, G.-S. Cho, N. Hamdani, R. Holewinski, S.-H. Jo, T. Danner, M. Zhang, P. P. Rainer, D. Bedja, J. A. Kirk, M. J. Ranek, W. R. Dostmann, C. Kwon, K. B. Margulies, J. E. Van Eyk, W. J. Paulus, E. Takimoto, D. A. Kass, Phosphodiesterase 9a controls nitric-oxide-independent cgmp and hypertrophic heart disease, Nature 519 (2015) 472–476.

[67] B. Borghese, J. Tost, M. de Surville, F. Busato, F. Letourneur, F. Mondon, D. Vaiman, C. Chapron, Identification of susceptibility genes for peritoneal, ovarian, and deep infiltrating endometriosis using a pooled sample-based genome-wide association study, BioMed Research International (2015).

[68] H. Lee, M.-c. A. Lin, H. I. Kornblum, D. M. Papazian, S. F. Nelson, Exome sequencing identifies de novo gain of function missense mutation in kcnd2

in identical twins with autism and seizures that slows potassium channel inactivation, Human Molecular Genetics 23 (2014) 3481.

[69] M. J. Perrin, A. Adler, S. Green, F. Al-Zoughool, P. Doroshenko, N. Orr, S. Uppal, J. S. Healey, D. Birnie, S. Sanatani, M. Gardner, J. Champagne, C. Simpson, K. Ahmad, M. P. van den Berg, V. Chauhan, P. H. Backx, J. P. van Tintelen, A. D. Krahn, M. H. Gollob, Evaluation of genes encoding for the transient outward current (ito) identifies the kcnd2 gene as a cause of j-wave syndrome associated with sudden cardiac deathclinical perspective, Circulation: Cardiovascular Genetics 7 (2014) 782–789.