Nam et al. BMC Bioinformatics (2019) 20:576
https://doi.org/10.1186/s12859-019-3106-9

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

# The translational network for metabolic disease – from protein interaction to disease co-occurrence

Yonghyun Nam[1†], Dong-gi Lee[1†], Sunjoo Bang[1], Ju Han Kim[2], Jae-Hoon Kim[1*] and Hyunjung Shin[1*]

## Abstract

**Background:** The recent advances in *human disease network* have provided insights into establishing the relationships between the genotypes and phenotypes of diseases. In spite of the great progress, it yet remains as *only a map of topologies between diseases,* but not being able to be a pragmatic diagnostic/prognostic tool in medicine. It can further evolve from a map to a *translational* tool if it equips with a function of scoring that measures the likelihoods of the association between diseases. Then, a physician, when practicing on a patient, can suggest several diseases that are highly likely to co-occur with a primary disease according to the scores. In this study, we propose a method of implementing '*n-of-1 utility*' (*n* potential diseases of *one* patient) to human disease network—*the translational disease network*.

**Results:** We first construct a *disease network* by introducing the notion of *walk* in graph theory to *protein-protein interaction network*, and then provide a *scoring algorithm* quantifying the likelihoods of *disease co-occurrence* given a primary disease. Metabolic diseases, that are highly prevalent but have found only a few associations in previous studies, are chosen as entries of the network.

**Conclusions:** The proposed method substantially increased *connectivity* between metabolic diseases and provided *scores* of *co-occurring diseases*. The increase in connectivity turned the disease network *info-richer*. The result lifted the AUC of random guessing up to 0.72 and appeared to be concordant with the existing literatures on *disease comorbidity*.

**Keywords:** Semi-supervised learning, Disease network, Comorbidity, Protein interaction, Disease scoring

## Background

The recent advances in human disease networks have provided insights into establishing the relationships between the genotypes and phenotypes of human diseases [1–3]. A disease (or disorder) is often thought of as resulting from rare mutations that trigger disruptions in underlying cellular functions. However, it is far from sufficient to define diseases solely by a mutation in a single gene because they are influenced by the totality of the intricate molecular connections between numerous cellular components [4–7]. A series of successful experiments developed in network biology have been beneficial for the progress of human disease network analysis [8, 9], which includes various types of molecular connections such as networks of gene co-expression, transcriptional regulations, protein interactions, metabolic pathways, and so on [10]. In [11], the authors provide a good review on the main features and the pros-cons of the existing methods for the disease-related biomolecular networks. Also, a comprehensive compilation of known disease-disease association can be found from many web services such as the DiseaseConnect (http://disease-connect.org) [12].

An initiative challenge for human disease networks was suggested by Goh et al. [13], which attempts to identify disease associations based on the genes that diseases share. Most diseases were grouped into several clusters; in particular, the cancer cluster is tightly interconnected owing to the many genes associated with multiple types of cancer. This has led to successful

* Correspondence: jayhoon@ajou.ac.kr; shin@ajou.ac.kr
†Yonghyun Nam and Dong-gi Lee contributed equally to this work.
[1]Department of Industrial Engineering, Ajou University, 206, World cup-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 16499, Republic of Korea
Full list of author information is available at the end of the article

research using various molecular networks. Zhang et al. (2011) proposed a disease network using the protein-protein interaction (PPI) network [14], motivated by studies indicating the genes that share similar or the same disease phenotypes tend to encode proteins that interact with each other [15, 16]. Indeed, the Hermansky–Pudlak syndrome [17] and Fanconi anemia [18] are known to be caused by mutations affecting different, but interacting, proteins. Lee et al. (2008) used the metabolic pathway network, and hypothesized that diseases are associated if they are linked to potentially correlated metabolic reactions [19]. Paik et al. (2014) proposed network-based disease–disease similarity analysis by focusing on topological similarity, suggesting disease-pathological symptoms through protein interactions [20]. A number of disease network studies have incorporated related methods [21–24].

Although our understanding of disease networks has expanded by virtue of the growth in theories and technical tools in the past, there is some room for improvement in the previous research. *First,* many works on disease networks have not identified tight associations for metabolic diseases. For cancer related diseases, a cancer cluster is successfully characterized and its associations to other diseases, including several diseases with a strong predisposition to cancer (such as Fanconi Anemia and Ataxia Telangiectasia), are well established because many genes are associated with multiple types of cancer (*TP53, KRAS, ERBB2, NF1,* etc.). However, for metabolic diseases, they are underrepresented, do not appear to form a distinct cluster, and have the fewest connections to other diseases. This result is attributed to the lack of information on genetic mutations associated with metabolic disease. But in practice, metabolic diseases may not really be so independent of one another, since certain metabolic diseases such as diabetes mellitus and hypoglycemia [25] or dyslipidemia and amyotrophic lateral sclerosis [26] often co-occur in the same individual and one can sometimes be considered a significant risk factor for the presence of the other. In the meantime, one can find another motivation for the research on metabolic disease network when consulting the report on prevalence statistics for diseases. A large proportion of people suffer from metabolic disruptions and the incidence rate is in trending upwards at the population level. During the period 2003 to 2012, metabolic syndrome prevalence in the United States was approximately, 18.0% among adults aged 20~39 years, 35.0% among adults aged 40~59 years, and 46.7% among adults aged 60 years and above [27]. With respect to diabetes in particular, 9.3% of the U.S. population has been diagnosed with diabetes, and 37% of U.S. adults aged 20 years or older were pre-diabetic in 2009–2012 [28]. On the other hand, the cancer incidence rate in 2011 was approximately 3.2% of the population and has declined year by year. The cancer death rate has also been declining by 1.5% per year for two decades. When comparing the statistics of metabolic diseases with those of cancer [29], one cannot lay less emphasis on the significance of metabolic diseases. Despite awareness of the importance, studies on the network analysis for metabolic diseases have not been yet well established. From those perspectives, it is of special interest to elucidate further the associations among metabolic diseases.

*Second*, disease networks have not been of benefit to medical research and practice yet, although they are poised to play a big role at the cellular level. The majority of the research on disease networks is still limited to developing a methodology to construct the network even when the approach makes use of sources of information from disease-gene associations, the interactions of proteins encoded by disease-related genes, or the metabolic pathways of diseases. We conjecture that it is because the research, in most cases, was issued and conducted by biologists pursuing purely scientific findings. In the perspective of physicians/clinicians/patients, however, this practice may be regarded as somewhat unkind since the results obtained from biology labs are too remote to be useful in the face of the reality of practicing medicine on patients. A doctor who is referencing a disease network when he/she treats a patient diagnosed with a particular disease may want to know if a co-occurrence is accidental or causal or if it increases the likelihood of the development of other diseases. It would be more convenient if the answer is given in the form of number, something like a score or a probability value, for disease co-occurrence. Unfortunately, most of the current disease networks do not provide quantified information such as disease co-occurrence scores, instead presenting a map of topologies between diseases. A recent study of Paik et al. (2014) provides the comorbidity information by comparing protein interaction-based disease network and medical reports-based disease network [20]. It gives a simple dichotomous information if two diseases are comorbid or not.

To address the limitations discussed above, we suggest a *disease network* model that provides quantified information, *scores* or *probabilities* for co-occurring diseases. To overcome sparse connectivity between metabolic diseases, we introduce the notion of *walk* of graph theory. The length of walk controls the range of protein-protein interactions that we use for identification of disease-disease associations. This idea is inspired by the definition of metabolic disease—the result of a genetic defect which frequently causes a metabolic enzyme to be non-expressed, inactive, or functionally compromised [30]. This implies that we may identify *latent* associations between metabolic diseases if we search deeper into the

PPI network. There may be hidden evidences that seemingly unrelated proteins interact somewhere in any of metabolic pathways. We then propose a method for *disease scoring*. Scores are calculated based on graph-based semi-supervised learning (SSL). The algorithm collects the *latent* information spread over the disease network to calculate scores. We validate the results of the proposed method with comorbidity literatures providing enrichment study for disease co-occurrence.

## Results

### Data for constructing disease networks

To construct a network of metabolic diseases, a list was obtained from MeSH in 2017. The National Library of Medicine has a controlled vocabulary thesaurus in the form of a taxonomy. When considering up to the second level of the taxonomy, there are 302 descriptors for metabolic diseases out of the 4663 listed diseases.

On the other hand, disease-protein relationship and protein-protein interactions data were obtained from the eight existing interaction databases: 53,480 disease-protein relationships between 2411 diseases and 7733 proteins, and 60,794 protein-protein interactions among 15,281 proteins, respectively. Based on Entrez gene and MeSH, we have curated all relational information related to metabolic diseases from multiple databases. The presence or absence of disease-protein relation was established to produce a binary disease vector. Table 1 summarizes the data sources. Of 302 metabolic diseases, only 181 have at least one disease-protein relationship, therefore, the size (the number of nodes) of our disease network was set to 181. (Additional file 1: Table A1 in Appendix A provides a full list of the 181 diseases.) The size of the PPI network is 15,281. The configuration of the two networks will be described as the upper and lower layers in Fig. 5a.

### Results for disease networks from the PPI network

The 181 metabolic diseases are represented as 15,281 dimensional vectors, where a disease vector is composed of binary attributes, each of which stands for presence ('1') or absence ('0') of the association with a particular protein as in Fig. 6. A disease network is built by calculating associations between disease vectors. Cosine distance is employed and is fed into the Gaussian similarity function (1): cosine distance is defined as $dist\ (\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 - (\ \boldsymbol{x}_i \cdot \boldsymbol{x}_j)/(\|\ \boldsymbol{x}_i\| \cdot \|\ \boldsymbol{x}_j\|)$, where $dist\ (x_i, x_j)$ is transformed to similarity using cosine distance between $x_i$ and $x_j$, '·' indicates the vector dot product, and $\|\boldsymbol{x}\|$ is the length of vector $\boldsymbol{x}$.

A disease vector for metabolic diseases is sparse since only a few disease-protein associations are known. This implies that the resulting disease network has sparse connectivity as well: the density of the matrix for disease-disease association is only 1.10%. See Fig. 1. Note that this network corresponds to $\boldsymbol{PPI}^{(0)}$ by the definitions in Methods section. In Fig. 1, the entry represents the counted number for disease-disease associations. The matrix connectivity increases as q increase. The heat-maps for disease-disease associations get denser as the step size $\boldsymbol{q}$ increases. The density reaches 26.87% in $\boldsymbol{PPI}^{(3)}$. The results show that we can widen the range of association of a disease (when there is only little information available) by applying the notion of $q$-step on the PPI network. And further, we can infer associations for most diseases and for rare diseases as well with the resulting disease network.

Figure 2 presents a toy demonstration for the proposed disease network vs. the existing one by Goh

**Table 1** Data for diseases, disease-protein relationships, and protein-protein interactions, and literature for comorbidity analysis: the number in parentheses indicates the amount of data originating from the respective sources. See also Additional file 1: Appendix A and C

| | Metabolic diseases | Disease-protein relationship | Protein-protein interaction | Comorbidity analysis |
|---|---|---|---|---|
| Data Sources | MeSH<br>The Medical Subject Headings<br>www.nlm.nih.gov/mesh/ | CTD (7624) ver. 2014/07/11<br>Comparative Toxicogenomics Database, www.ctdbase.org/<br>GAD (34,773) ver.2013/07/16<br>Genetic Association Database, www.geneticassociationdb.nih.gov/<br>OMIM (4078) ver. 2014/03/10<br>Online Mendelian Inheritance in Man, www.omim.org/<br>PharmGKB (6610) ver. 2015/07/20<br>The Pharmacogenomics Knowledge Base, www.pharmgkb.org/<br>TTD (395) ver. 2013/07/04<br>Therapeutic Target Database, www.bidd.nus.edu.sg/group/cjttd/ | DIP (773) ver. 2014/10/10<br>Database of Interacting Proteins, www.dip.doebi.ucla.edu/dip/Main.cgi<br>Entrez Gene (58,778)<br>ver. 2014/07/20<br>www.jura.wi.mit.edu/entrez_gene/<br>MINT (736) ver. 2014/07/08<br>Molecular Interaction Database, www.mint.bio.uniroma2.it/mint/<br>PharmGKB (507)<br>ver. 2014/07/20<br>The Pharmacogenomics Knowledge Base, www.pharmgkb.org/ | PubMed Literature<br>US National Library of Medicine<br>National Institutes of Health |
| Number of Data | 181 out of 302 metabolic diseases | 53,480 relations between 2411 diseases and 7733 proteins | 60,794 interactions of 15,281 proteins | 62 pairs of 55 diseases |

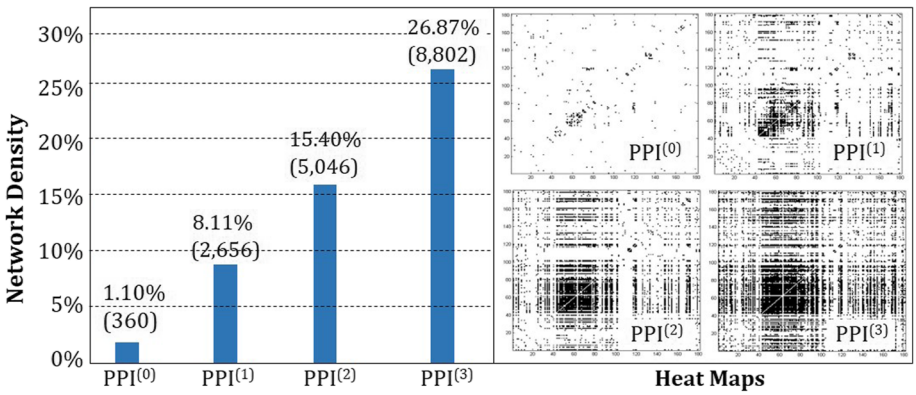


**Fig. 1** Changes in network density by a q-step walk on the PPI network: The amount of connections for disease-disease association increases as the step size q increases. With rich connections among diseases, the chances of having inferences for most diseases, including rare diseases, are increased because it enables us to use the information propagated from other diseases through the connections

et al.'s. The disease network is composed of eight metabolic diseases such as hyperlipoproteinemia, homocystinuria, maple syrup urine disease, pyruvate dehydrogenase deficiency, lipodystrophy, insulin resistance, Fanconi syndrome, and congenital hyperinsulinism. The dotted line indicates the edge by Goh et al.'s: there is only a single edge connecting two diseases and the rest are disconnected. In contrast, in $PPI^{(integrated)}$ disease network (a network piling up $PPI^{(0)}$ up to $PPI^{(3)}$), the diseases are all connected, and the connection between the nodes is large with various connection strengths. The width of an edge is proportional to the number of shared proteins. These are shown in grey solid lines in Fig. 2. The comparison results show that we are now able to make inferences about most diseases (including rare diseases) based on these richer connections of the disease network. Additional file 1: Fig. B1 in Appendix B provides a full network comparison. (The Matlab source code is available in Additional file 2: 'NetworkConstruction.m'.)

## Results for disease scoring

### *Enrichment with comorbidity study*

To validate the proposed method as reliable in real medical/clinical practice, we adopted data from the literature of a comorbidity study as an independent source of information for this test. Comorbidity measures the presence of one or more additional diseases (or disorders) co-occurring with a primary disease or the effects of such additional diseases. The sources of information, used by the researchers and scientists working on the question of comorbidity, are case histories [31, 32], hospital records of patients [33] and other medical documentation kept by family doctors, or insurance companies [34]. Therefore, data in the literature from comorbidity studies are mainly based on the clinical experience and qualifications of the physicians carrying out clinical, instrumental, and laboratory-confirmed diagnoses. In total, 62 pairs of 55 comorbid diseases

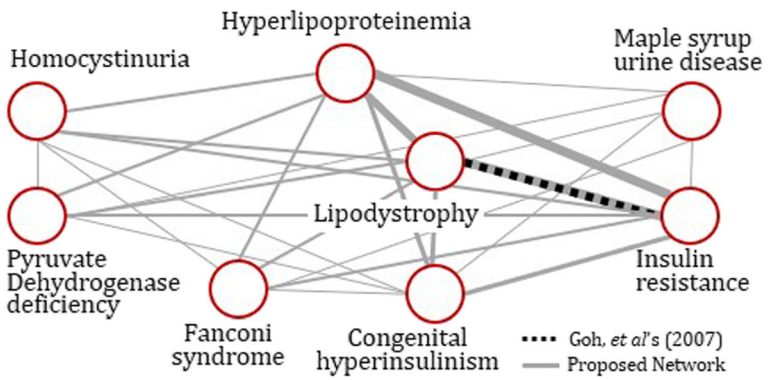


**Fig. 2** Disease network comparison: A small subset of diseases belonging to the metabolic diseases category in MeSH is selected. The nodes in the graph represent the eight selected diseases, and the width of an edge indicates connection strength proportional to the number of shared proteins between two diseases. In Goh et al.'s approach, there is only a single edge connecting two disease nodes, while the remaining six are disconnected. In contrast, in $PPI^{(0\sim q)}$, all eight disease nodes are connected, and each has a high degree of node connectivity with various connection strengths

were obtained from a literature survey. See the last column of Table 1. The full list of the comorbidity literature used in this study is provided in Table C1, and the references are provided in Additional file 1: Appendix C.

### *Comparative experiments for scoring*

The proposed SSL based scoring method (in Methods section) was applied to $\boldsymbol{PPI}^{(q)}$ 's, hypothesizing that our extended disease networks will provide improved scoring as compared to the existing disease network produced by Goh et al. Note that Goh et al.'s approach coincides with that of our $\boldsymbol{PPI}^{(0)}$ network. In order to obtain a random control as a reference point, we created a randomized network; we shuffled disease-disease associations while keeping both the number of connected nodes and the degree of a node the same as those of $\boldsymbol{PPI}^{(integrated)}$, and projected it onto a disease network. The loss-smoothness tradeoff parameter $\mu$ in (5) was set to a large number ($\mu = 100$). From the published comorbidity studies, 55 out of 181 diseases were able to obtain labels in our SSL scoring model. For instance, by setting the label of a disease on ' $y_l = 1$ ', while keeping unchanged the labels of the remaining 180 diseases as ' $y_u = 0$ ', we obtained the predicted score for identifying the comorbid diseases with the given disease. The experiment was repeated 55 times and the performance was measured by AUC (the area under the receiver operating characteristic curve) [35]. The AUC was obtained by comparing predictive value $\boldsymbol{f}$ in (5) and PubMed Literatures: presence ('1') or absence ('0') of PubMed literatures. It is used as a standard for disease association and comorbidity disease.

The AUC performances for the disease networks are summarized in Fig. 3. The randomized disease network produces an AUC of 0.5, which is only equivalent to the performance by random guessing. However, $\boldsymbol{PPI}^{(0)}$ lifts the performance up to 0.69. Thereafter, the performance of the disease network continuously increases as the step size $q$ increases up to $q = 3$: $\boldsymbol{PPI}^{(0)} < \boldsymbol{PPI}^{(1)} < \boldsymbol{PPI}^{(2)} < \boldsymbol{PPI}^{(3)}$. Given these results, it is plausible that the more connected a disease network, the higher the AUC performance will be. After $q = 3$, however, the performance stayed unchanged till $q = 6$, then begins to degrade. We conjectured that the complexity of $\boldsymbol{PPI}^{(3)}$, in terms of

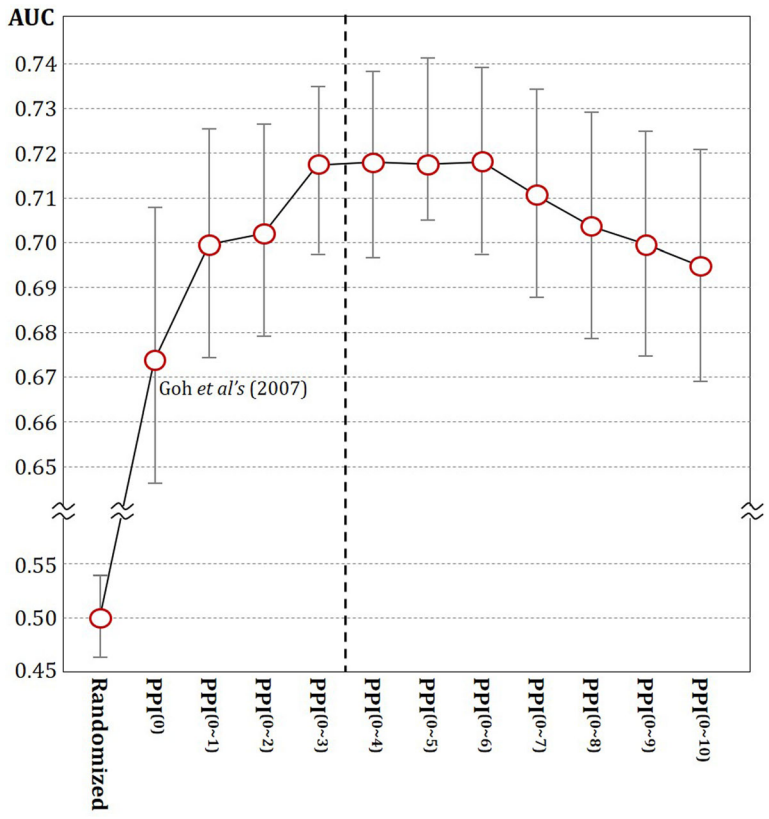


**Fig. 3** AUC comparison from $q = 0$ to 10 for integrated networks: $PPI^{(0\sim q)}$. The experiment was repeated 55 times, and the average AUC value with the standard deviation is presented as a circle with an error bar. $PPI^{(0)}$, $PPI^{(1)}$, $PPI^{(2)}$, and $PPI^{(3)}$ are individual disease networks constructed as described in Methods section. $PPI^{(0\sim q)}$ s are integrated networks from Eq. (6). $PPI^{(0)}$ corresponds to the existing disease network suggested by Goh et al. [21]. A randomized network is added to our experiment to obtain a reference performance. The best performance was achieved by $PPI^{(0\sim 3)}$, and *p*-values of the pairwise t-tests are shown in the bottom of the plot

network connectivity, is enough to draw most of information in data, therefore more complication may not be needed. Additional file 1: Fig. D1 in Appendix D shows performance of the individual networks. On the other hand, it is interesting to observe that combining $\boldsymbol{PPI}^{(q)}$ to $\boldsymbol{PPI}^{(q+1)}$ ($q = 0, 1, 2$) further increases the AUC performance. $\boldsymbol{PPI}^{(0\sim1)}$ is an integrated network piling up $\boldsymbol{PPI}^{(0)}$ onto $\boldsymbol{PPI}^{(1)}$, and $\boldsymbol{PPI}^{(0\sim2)}$ is similarly constructed. This pattern of improvement ends with $\boldsymbol{PPI}^{(0\sim3)}$, resulting in an AUC of 0.72. The further integration did not make significant improvement in performance, so it was decided as our best network. The $p$-values of the pairwise $t$-test (Additional file 1: Table D1 in Appendix D) demonstrate the statistical significance of the improved performance of $\boldsymbol{PPI}^{(0\sim3)}$ over the others. Additional experiments were conducted to verify the performance instead of leave-one-out. 5-fold cross validation was performed for three disease groups: metabolic diseases, neoplasms, and nervous system diseases. The experiment was repeated ten times, and for each of them, 5CV was conducted after random permutation of data. The results were summarized in an Additional file 1: Appendix E.

## Discussion

### Implication of the probabilities of the associated diseases

Figure 4 depicts a typical example of the proposed scoring results with the probabilistic transfer function (6). Diabetes mellitus type II (T2DM) was set as the labeled (target) disease, and then the probability values for association with the remaining 180 diseases were obtained from the integrated network, $\boldsymbol{PPI}^{(0\sim3)}$. The solid line in the figure stands for the probability values of the 180 diseases. The open circles on the line correspond to the diseases comorbid with T2DM evidenced by the literature. The below shows the clinical implications for some of the marked comorbid diseases observed in the literature. More evidences can be found in Additional file 1: Appendix F.

#### *High in insulin resistance[a] and hyperinsulinism[b]*

T2DM is preceded or accompanied by an elevated adiposity that causes insulin resistance [36]. Insulin resistance causes hyperinsulinemia to maintain normal glucose levels. When the pancreas cannot sustain hyperinsulinemia to overcome insulin resistance, pre-diabetes or T2DM ensues [37].

#### *High in lactic acidosis[c]*

Lactic acidosis is caused by accumulation of lactic acid more rapidly than it can be metabolized. It may occur spontaneously or in association with diseases such as T2DM, leukemia, or liver failure [38].

#### *High in diabetic ketoacidosis[d] vs. low in acidosis[e]*

Presence of Diabetic ketoacidosis has been increasingly recognized in patients with T2DM, and a newer entity called ketosis-prone diabetes is also commonly recognized [39, 40]. Diabetic ketoacidosis in patients with T2DM tends to present with a less severe acidosis and patients are more likely to have normal potassium levels [41–43].

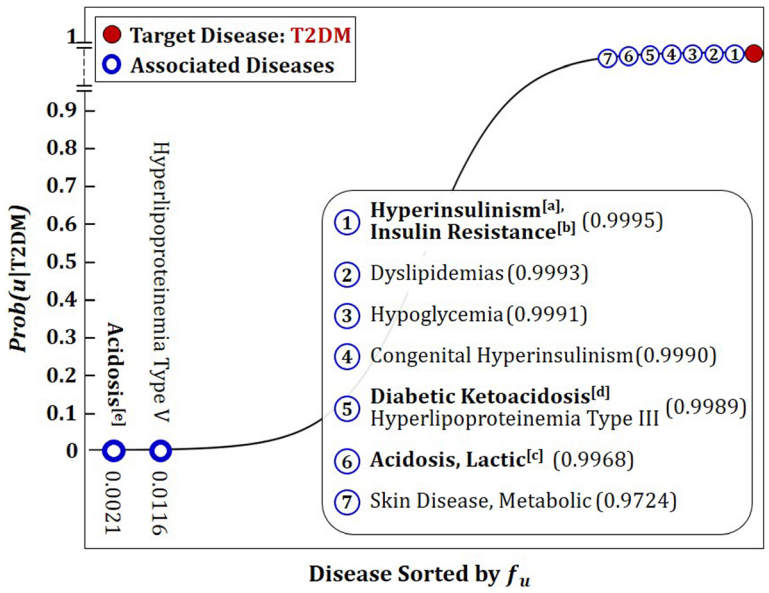


**Fig. 4** Probabilities of the diseases associated with T2DM: The solid line represents the probability values of the 180 diseases. Shown on the line with open circles are the locations of 13 diseases comorbid with diabetes mellitus type II. The probability values support the knowledge based on real medical practice and vice versa

## Conclusion

In this study, we proposed two novel methods: the method of disease network construction by extracting latent information from different depths of layers of the PPI network and the method of disease scoring based on SSL by collecting the latent information spread over the network. To examine whether the proposed method for disease co-occurrence provides predictions within reasonable bounds in practice, we investigated pairs of comorbid diseases that were reported in the literature and compared them with the obtained scores. The result was promising; the scoring results appear concordant with conventional comorbidity studies.

There are some noteworthy features of the present study: (a) Despite great progress in research on disease networks, there are still barriers for physicians to use it in practice: a disease network has been little more than a map of topologies between diseases. It is inconvenient to deduce the co-occurrence of the associated diseases, and they do not have enough confidence to put it into action when practicing a patient. From this point of view, this study suggests a streamlined methodology with biologically driven knowledge—the protein-protein interaction data— and the scores for disease co-occurrences. This will eventually assist physicians in adopting smarter strategies earlier to aid them in tackling the numerous intricacies inherent to the treatment of diseases.

(b) The proposed method of constructing a disease network is an unprecedented and systematic approach. Previous work identifies disease-disease associations based on the gene that diseases share [13], based on the gene encoding the protein that interacts with the protein encoded by other disease genes [14], or based on shared metabolites and correlated metabolic reactions [19]. The present study provides a framework embracing the previous work: Goh et al. (2007)'s is analogous to our $\boldsymbol{PPI}^{(\mathbf{0})}$, Zhang et al. (2011)'s to $\boldsymbol{PPI}^{(\mathbf{0\sim1})}$, Lee et al. (2008)'s to $\boldsymbol{PPI}^{(q)}$ where $q$ is arbitrary depending on the size of metabolic pathway, and Paik et al. (2014)'s to $\boldsymbol{PPI}^{(q)}$ where $q$ is 0 and 1. With our methodology, it becomes simple to expand or shrink the reference ranges of the PPI network—just adjust the step size $q$ of walk! With the ongoing growth in study for the completeness of the PPI network such as [44] and the references therein, we expect that the proposed methodology on how to construct a disease network from PPI will further power the disease mechanism research.

(c) An algorithm for disease scoring must be able to deal with the circumstance that only limited labeled data are provided because a patient will provide only a few pieces of information on one or two diseases that he/she has contracted. SSL can handle these situations, and this trait has inspired us to develop a scoring algorithm based on SSL.

In this study, we have only focused on the connectivity of diseases via genes without considering they are essential or redundant. We could not disregard a gene even if it is regarded as redundant because it can be used as a link to reach other diseases. We provided a typical case about how a disease is linked to other diseases, and therein, which gene turns out to be important (essential) for the connection (See Additional file 1: Appendix G). We note that the current study is motivated by and developed for metabolic diseases that are known to be the most disconnected class for most human disease networks. Extension of this methodology to the whole set of human diseases should be attempted next. We added some preliminary results for other categories of disease to Additional file 1: Appendix H. Another extension may be developed by incorporating diverse data sources to our method. Particularly, if updating the translational disease network with cleaner and higher quality of data [45], we expect that it will be upgraded to be more reliable and accurate.

## Methods

The translational disease networks consist of two methods: (a) *a method of constructing a disease network* based on protein-protein interaction data and (b) *a network-based scoring method* for calculating the probabilities of disease co-occurrence when a specific disease is given. The resulting disease network from the first becomes the base on which the second works. See (a) and (b) in Fig. 5.

### Disease networks by q-step walks on the PPI network

The disease network is a graph, $G(V, W)$, that represents the associations between pairs of diseases by assigning a weight to the edge connecting two diseases. In the network, the node $V$ denotes diseases and the weight $W$ denotes similarity between the sequences of proteins that two diseases commonly share. In the proposed method, the notion of *commonly shared proteins* is expanded by applying the *walk* (or *path*) of graph theory to the PPI network. The step size of walk on the PPI network determines the number of shared proteins. On a graph, a '*walk*' starting at node $v_A$ and ending at node $v_B$, is represented as $(v_A \rightarrow v_1 \rightarrow \ldots \rightarrow v_n \rightarrow v_B)$. The edges connect the successive nodes in a walk. Let us define a '*q-step walk*' as a walk of length $q$, which travels $q$ edges departing from $v_A$ for $v_B$. In a conventional approach for constructing a disease network, the two diseases are defined as *associated* only if they are known to share same proteins. In Fig. 5a, $D_B$ and $D_C$ share protein $P_1$, therefore they are considered as associated. Figure 6a rephrases this in disease-protein vector representation, and we see that there exists no other disease association than a single one between $D_B$ and $D_C$. In terms of $q$-step walk, it corresponds to the case of $q = 0$.

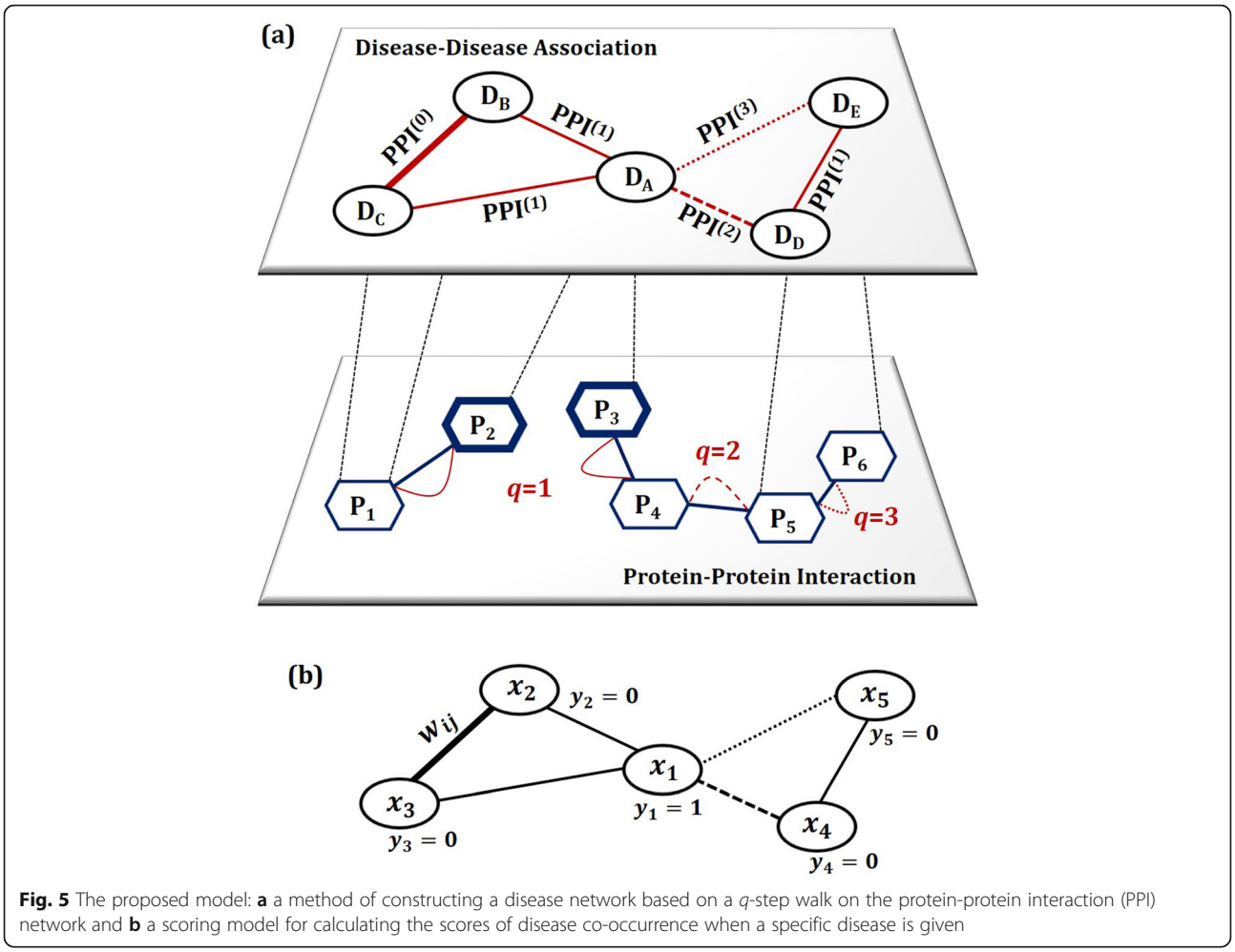


**Fig. 5** The proposed model: **a** a method of constructing a disease network based on a *q*-step walk on the protein-protein interaction (PPI) network and **b** a scoring model for calculating the scores of disease co-occurrence when a specific disease is given

However, if the notion of *q*-step walk is applied to the PPI network, disease associations can be further expanded. Consider $D_A$ in Fig. 5a which is known to be related to two proteins $P_2$ and $P_3$. And it is not associated with any of diseases in 0-step walk. An 1-step walk departing from those proteins reaches $P_1$ and $P_4$, respectively, (see Fig. 5a), and produces a different disease-protein vector (see Fig. 6b). Then, it now can be associated with $D_B$ and $D_C$. Similarly, a 2-step walk ( $P_3 \rightarrow P_4 \rightarrow P_5$) or a 3-step walk ($P_3 \rightarrow P_4 \rightarrow P_5 \rightarrow P_6$) makes expanded associations with other diseases, which are described in Fig. 6c and d.

Applying different *q*-step walk to the PPI network, a disease network can be differently constructed, and is denoted as $\boldsymbol{PPI}^{(q)}$. The following lists general definitions of $\boldsymbol{PPI}^{(q)}$ 's where $q = 0, 1, 2, 3$. The disease network by *q*-step walks on PPI network is a graph $PPI^{(q)} = G(V, W^{(q)})$, where $W^{(q)}$ is the similarity after applying *q*-step walks. (Note that the initial network, $\boldsymbol{PPI}^{(0)}$, is original disease network $G(V, W)$). Consider two diseases, Disease I and Disease II, assuming that the former is known to be associated with protein $P_I$ and the latter with $P_{II}$. The step-by-step process of *q*-step walk is described in detail in Additional file 1: Appendix I.

- $\boldsymbol{PPI}^{(0)}$: The two diseases are defined as *associated* only if $P_I$ and $P_{II}$ are identical ($P_I \equiv P_{II}$)—the disease network by *0-step walk*.
- $\boldsymbol{PPI}^{(1)}$: An association between the two diseases is defined when the proteins are known to *interact* ($P_I \sim P_{II}$ where '~' stands for interaction)— the disease network by *1-step walk*.
- $\boldsymbol{PPI}^{(2)}$: Let us introduce an extra protein $P_{III}$ which interacts with the two disease proteins ($P_{III} \sim P_I$, $P_{III} \sim P_{II}$). Through bypassing the medium protein, $P_I \rightarrow P_{III} \rightarrow P_{II}$ and vice versa, the two diseases are related—the disease network by *2-step walk*.
- $\boldsymbol{PPI}^{(3)}$: Similarly, a *3-step walk* on the PPI network, $P_I \rightarrow P_{III} \rightarrow P_{IV} \rightarrow P_{II}$, can bridge the two diseases via two medium proteins, $P_{III}$ and $P_{IV}$—the disease network by *3-step walk*.

One may further develop $\boldsymbol{PPI}^{(q)}$ 's by increasing the walk length *q*. As *q* increases, more associations can be

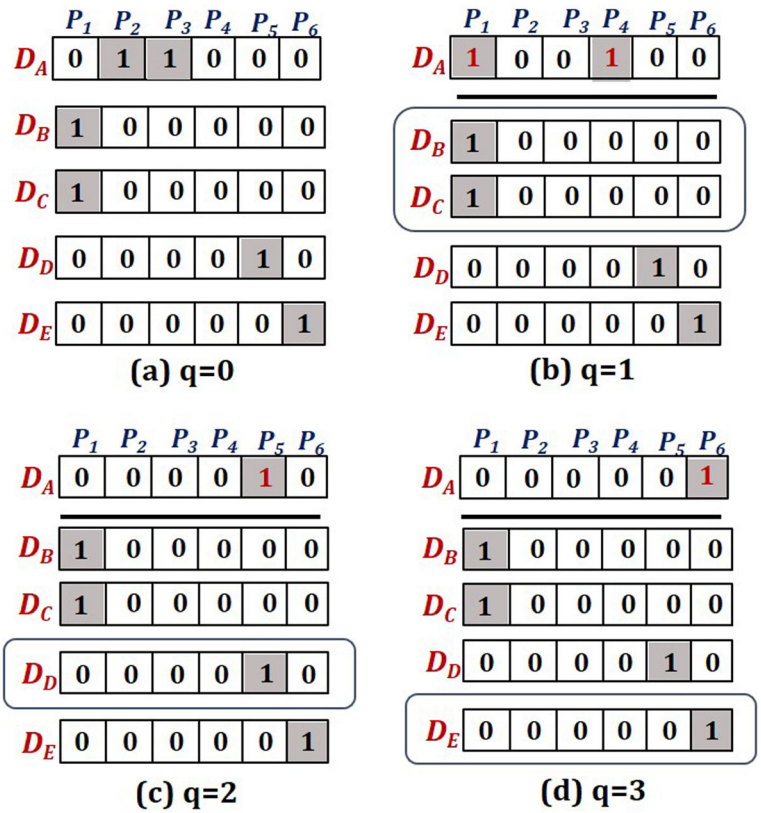


**Fig. 6** Associated diseases of $D_A$ by $q$-step walks on the PPI network in the case of **a** $q$=0, **b** $q$=1, **c** $q$=2 and **d** $q$=3

found among diseases, and the density of the disease network increases. It is conjectured that a network of a larger step size is relatively less informative because of loss of information while touring around the network. On the contrary, a small step size $q$ provides a good quality of associations between diseases but it may result in a disconnected disease network where many of the diseases remain isolated (or dangled), and is described in this paper as a drawback of the status quo approaches [13, 14]. Note that disease associations based on the above definitions are reciprocal; hence, a resulting disease network, $\boldsymbol{PPI}^{(q)}$, is symmetric.

### A scoring model for co-occurring diseases

Once a disease network is obtained, a person who has caught/contracted a particular disease may wish to be informed regarding *how likely he/she is to be exposed to other diseases*. Hypothesizing that *at least* one disease is known, the proposed *scoring* method calculates probabilities for the associated diseases.

#### *Scoring algorithm*

Let us define *disease scoring* as a function that quantifies the degree of commitment of the associated diseases when one or a few diseases are given. To embody scoring in a disease network, $\boldsymbol{PPI}^{(q)}$, we employ the conventional settings for the graph-based semi-supervised learning (graph-based SSL) classification and modify it to be suitable for our scoring problem. SSL has attracted the interests of many researchers in areas where labeled data are a few but unlabeled data are abundant [5, 46–52]. And it has been reported that SSL successfully improves classification performance by supporting classifiers with unlabeled data [5]. The motivation of SSL is appealing for our problem because it will be typical for a patient to have contracted one or a few diseases (labeled data) but not the rest of the diseases (unlabeled data). To implement an SSL based scoring algorithm one has to perceive the difference between classification and scoring. In a (binary) classification problem, the labels given to a classifier are binary (+ 1 or – 1), and the resulting prediction is made by the way that each of the unlabeled data are assigned to either one class (+ 1) or the other (– 1). On the contrary, in a scoring problem, unary labels (+ 1) are given to a scorer, and the resulting predictions are scores that prioritize the unlabeled data for the given labels. In the proposed method, it also outputs the corresponding probability values. Figure 5b schematically describes our graph-based SSL scoring method, and the following paragraphs explain the details.

Consider a graph $G = (V, W)$ with node $V$ corresponding to the $n(=n_l + n_u)$ data points from labeled set $S_L = \{(x_i, y_i)_{i=1}^{n_l}\}$ and unlabeled set $S_U = \{(x_i)_{i=n_l+1}^{n}\}$. In the proposed SSL based scoring problem, the $n_l$ nodes are set to a unary label $\boldsymbol{y}_l \in \{1\}$ while the unlabeled $n_u$ nodes are set to zero ($\boldsymbol{y}_u \in \{0\}$). The task is to assign scores $\boldsymbol{f}_u^T = (f_{n_l+1}, \ldots, f_n)^T$ on nodes $V_U$. To compute a real-valued scoring function $f: V \rightarrow \mathbb{R}$ on $G$, one strategy is to let the label information propagate to the unlabeled nodes through edges $W$. The edge weight $w_{ij}$ between the two nodes $x_i$ and $x_j$ can take a value of 0 or 1 in the simplest case. Usually, $dist\ (x_i, x_j)$ between $x_i$ and $x_j$ is transformed to similarity using Gaussian

$$w_{ij} = \begin{cases} \exp^{-dist\ (x_i, x_j)/\sigma^2} & if\ i \sim j \\ 0 & otherwise \end{cases} \quad (1)$$

where $i \sim j$ indicates that the two nodes are connected. The connection strength is encoded in $w_{ij}$ of a similarity matrix $W$, and a large value of $w_{ij}$ represents more similarity between the two nodes. Assuming that $f_i$ should be close to the given label $y_i$ in labeled nodes—loss condition, and overall, $f_i$ should not be too different from the $f_j$ of adjacent nodes—smoothness condition, one can obtain $\boldsymbol{f}$ by minimizing the following quadratic functions:

$$\min_{\boldsymbol{f}} H(\boldsymbol{f}) = (\boldsymbol{f} - \boldsymbol{y})^T (\boldsymbol{f} - \boldsymbol{y}) + \mu \boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f} \quad (2)$$

where $\boldsymbol{y} = [y_1, \ldots, y_{n_l}, 0, \ldots, 0]^T$. The matrix $\boldsymbol{L}$ known as the graph Laplacian matrix, is defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ where $\boldsymbol{D} = diag\ (d_i)$, $d_i = \sum_j w_{ij}$. The user parameter $\mu$ trades off loss (the first term of $H(.)$) and smoothness (the second term). The solution of this problem becomes

$$\boldsymbol{f} = (\boldsymbol{I} + \mu \boldsymbol{L})^{-1} \boldsymbol{y} \quad (3)$$

One may compute the scores for the unlabeled nodes explicitly in a block-wise representation of the similarity matrix $W = [W_{ll}\ W_{lu} \mid W_{ul}\ W_{uu}]$. Let us represent (3) as a block structure and rearrange $\boldsymbol{y}$ to the left side of the equality,

$$\begin{bmatrix} \boldsymbol{y}_l \\ \boldsymbol{y}_u \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} + \mu(D_{ll} - W_{ll}) & -\mu W_{lu} \\ -\mu W_{ul} & \boldsymbol{I} + \mu(D_{uu} - W_{uu}) \end{bmatrix} \times \begin{bmatrix} \boldsymbol{f}_l \\ \boldsymbol{f}_u \end{bmatrix}. \quad (4)$$

Then, one can simplify (4) by substituting $\boldsymbol{f}_l = \boldsymbol{y}_l$ and $\boldsymbol{y}_u = \boldsymbol{0}$, and by writing it in terms of $\boldsymbol{f}_u$, we obtain the scores for the unlabeled nodes,

$$\boldsymbol{f}_u = \mu\{\boldsymbol{I} + \mu(D_{uu} - W_{uu})\}^{-1} W_{ul} \boldsymbol{y}_l. \quad (5)$$

On a network of diseases, by setting $\boldsymbol{y}_l = 1$ to the nodes for the contracted diseases, one can obtain scores $\boldsymbol{f}_u$ for the rest of diseases from (5).

### *Probability calculation*

After obtaining scores from a disease network, the next step is to transform the scores to probability values of disease co-occurrence. The resulting scores $\boldsymbol{f}_u$ from (5) is unique and satisfies $0 < \boldsymbol{f}_u < 1$ . After normalizing the overall $\boldsymbol{f}_u$ 's, the scores can be associated with probability values as below

$$Prob(u|l) = \frac{1}{1 + exp^{-f_u/\sigma_f}} \quad (6)$$

where $\sigma_f$ is a scale parameter. Given primary diseases ($l$), the output value of (6) measures *probability values* of *disease co-occurrence* for other diseases ($u$). One can refer to the values when attempting to figure out the secondary or tertiary diseases to the given one.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1186/s12859-019-3106-9.

**Additional file 1.** Supplementary materials for the proposed method and results.

**Additional file 2.** The matlab source code for disease network construction from PPI network.

### Abbreviations

CTD: Comparative Toxicogenomics Database; DIP: Database of Interacting Proteins; GAD: Genetic Association Database; MeSH: The Medical Subject Headings; MINT: Molecular Interaction Database; OMIM: Online Mendelian Inheritance in Man; PharmGKB: The Pharmacogenomics Knowledge Base; TTD: Therapeutic Target Database

### Authors' contributions

HJS designed the idea, wrote/revised the manuscript and supervised the study process. YHN analyzed the data, implemented the system, validated the results, and wrote/revised the manuscript. JaHK analyzed data and validated the results. DGL, SJB, JuHK implemented and validated the results. And all authors read and approved the final manuscript.

### Funding

HJS would like to gratefully acknowledge supported from the National Research Foundation of Korea (NRF) grant funded by the Korea government (MOE) (NRF-2018R1D1A1B07043524), and the Ajou University research fund. This study was also provided with biospecimens and data from the biobank of Chronic Cerebrovascular Disease consortium. The consortium was supported and funded by the Korea Centers for Disease Control and Prevention (#4845–303). JHK[1] would like to gratefully acknowledge supported from the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2017R1A2B1009709). The funders had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

### Availability of data and materials

The data can be found in PharmDB (http://pharmdb.org/). PharmDB is a tripartite pharmacological network database of human diseases, drugs, and

proteins which compiles and integrates nine existing interaction databases (Access date: 2016. 11. 03).

#### Ethics approval and consent to participate
Not applicable.

#### Consent for publication
Not applicable.

#### Competing interests
The authors declare that they have no competing interests.

#### Author details
[1]Department of Industrial Engineering, Ajou University, 206, World cup-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 16499, Republic of Korea. [2]Seoul National University Biomedical Informatics (SNUBI), Division of Biomedical Informatics, Seoul National University College of Medicine, 103, Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea.

Received: 21 August 2018 Accepted: 20 September 2019
Published online: 13 November 2019

#### References
1. Davis DA, Chawla NV. Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. PLoS One. 2011;6(7):e22670.
2. Hidalgo CA, Blumm N, Barabási A-L, Christakis NA. A dynamic network approach for the study of human phenotypes. PLoS Comput Biol. 2009;5(4):e1000353.
3. Streelman JT, Kocher TD. From phenotype to genotype. Evol Dev. 2000;2(3):166–73.
4. Argmann CA, Chambon P, Auwerx J. Mouse phenogenomics: the fast track to "systems metabolism". Cell Metab. 2005;2(6):349–60.
5. Kim J, Shin H. Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. J Am Med Inform Assoc. 2013;20(4):613–8.
6. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet J-P, Subramanian A, Ross KN. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313(5795):1929–35.
7. Loscalzo J, Kohane I, Barabasi AL. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. Mol Syst Biol. 2007;3(1):124.
8. Jothi R, Kann MG, Przytycka TM. Predicting protein–protein interaction by searching evolutionary tree automorphism space. Bioinformatics. 2005;21(suppl_1):i241–50.
9. Kelley R, Ideker T. Systematic interpretation of genetic interactions using protein networks. Nat Biotechnol. 2005;23(5):561.
10. Park S, Yang J-S, Kim J, Shin Y-E, Hwang J, Park J, Jang SK, Kim S. Evolutionary history of human disease genes reveals phenotypic connections and comorbidity among genetic diseases. Sci Rep. 2012;2:757.
11. Liu Z-P, Wang Y, Zhang X-S, Chen L-N. Network-based analysis of complex diseases. IET Syst Biol. 2012;6(1):22–33.
12. Liu C-C, Tseng Y-T, Li W, Wu C-Y, Mayzus I, Rzhetsky A, Sun F, Waterman M, Chen JJ, Chaudhary PM. DiseaseConnect: a comprehensive web server for mechanism-based disease–disease connections. Nucleic Acids Res. 2014;42(W1):W137–46.
13. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. Proc Natl Acad Sci. 2007;104(21):8685–90.
14. Zhang X, Zhang R, Jiang Y, Sun P, Tang G, Wang X, Lv H, Li X. The expanded human disease network combining protein–protein interaction information. Eur J Hum Genet. 2011;19(7):783.
15. Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tümer Z, Pociot F, Tommerup N. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol. 2007;25(3):309.
16. Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein–protein interactions. J Med Genet. 2006;43(8):691–8.
17. Di Pietro SM, Dell'Angelica EC. The cell biology of Hermansky–Pudlak syndrome: recent advances. Traffic. 2005;6(7):525–33.
18. Macé G, Bogliolo M, Guervilly J-H, du Villard JAD, Rosselli F. 3R coordination by Fanconi anemia proteins. Biochimie. 2005;87(7):647–58.
19. Lee D-S, Park J, Kay K, Christakis N, Oltvai Z, Barabási A-L. The implications of human metabolic network topology for disease comorbidity. Proc Natl Acad Sci. 2008;105(29):9880–5.
20. Paik H, Heo H-S, Ban H-J, Cho SB. Unraveling human protein interaction networks underlying co-occurrences of diseases and pathological conditions. J Transl Med. 2014;12(1):99.
21. Hu G, Agarwal P. Human disease-drug network based on genomic expression profiles. PLoS One. 2009;4(8):e6536.
22. Matias Rodrigues JF, Wagner A. Evolutionary plasticity and innovations in complex metabolic reaction networks. PLoS Comput Biol. 2009;5(12):e1000613.
23. Yıldırım MA, Goh K-I, Cusick ME, Barabási A-L, Vidal M. Drug—target network. Nat Biotechnol. 2007;25(10):1119.
24. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási A-L. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015;347(6224):1257601.
25. Lopez JM, Annunziata K, Bailey RA, Rupnow MF, Morisky DE. Impact of hypoglycemia on patients with type 2 diabetes mellitus and their quality of life, work productivity, and medication adherence. Patient Prefer Adherence. 2014;8:683.
26. Dupuis L, Corcia P, Fergani A, De Aguilar J-LG, Bonnefont-Rousselot D, Bittar R, Seilhean D, Hauw J-J, Lacomblez L, Loeffler J-P. Dyslipidemia is a protective factor in amyotrophic lateral sclerosis. Neurology. 2008;70(13):1004–9.
27. Yankey B, Rothenberg R, Strasser S, White K, Okosun I. Relationship between years of marijuana use and the four main diagnostic criteria for metabolic syndrome among United States adults. J Addict Res Ther S. 2017;11:2.
28. Control CfD, Prevention. National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014. Atlanta: US Department of Health and Human Services 2014; 2014.
29. Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. CA Cancer J Clin. 2014;64(1):9–29.
30. Curley RW. Retinoid chemistry: synthesis and application for metabolic disease. Biochim Biophys Acta (BBA)-Mol Cell Biol Lipids. 2012;1821(1):3–9.
31. Fuchs Z, Blumstein T, Novikov I, Walter-Ginzburg A, Lyanders M, Gindin J, Habot B, Modan B. Morbidity, comorbidity, and their association with disability among community-dwelling oldest-old in Israel. J Gerontol Ser A Biol Med Sci. 1998;53(6):M447–55.
32. Hoffman C, Rice D, Sung H-Y. Persons with chronic conditions: their prevalence and costs. JAMA. 1996;276(18):1473–9.
33. Daveluy C, Québec S, Québec Idlsd. Enquête sociale et de santé 1998: cahier technique et méthodologique. Québec: Institut de la statistique du Québec; 2001.
34. Wolff JL, Starfield B, Anderson G. Prevalence, expenditures, and complications of multiple chronic conditions in the elderly. Arch Intern Med. 2002;162(20):2269–76.
35. Swets JA. Signal detection theory and ROC analysis in psychology and diagnostics: collected papers. New York: Psychology Press; 2014.
36. Vinson GP. Angiotensin II, corticosteroids, type II diabetes and the metabolic syndrome. Med Hypotheses. 2007;68(6):1200–7.
37. Festa A, Williams K, D'Agostino R, Wagenknecht LE, Haffner SM. The natural course of β-cell function in nondiabetic and diabetic individuals: the insulin resistance atherosclerosis study. Diabetes. 2006;55(4):1114–20.
38. Fimognari FL, Pastorelli R, Incalzi RA. Phenformin-induced lactic acidosis in an older diabetic patient: a recurrent drama (phenformin and lactic acidosis). Diabetes Care. 2006;29(4):950–1.
39. Barski L, Nevzorov R, Jotkowitz A, Rabaev E, Zektser M, Zeller L, Shleyfer E, Harman-Boehm I, Almog Y. Comparison of diabetic ketoacidosis in patients with type-1 and type-2 diabetes mellitus. Am J Med Sci. 2013;345(4):326–30.
40. Puttanna A, Padinjakara R. Diabetic ketoacidosis in type 2 diabetes mellitus. Pract Diabetes. 2014;31(4):155–8.
41. Balasubramanyam A, Zern JW, Hyman DJ, Pavlik V. New profiles of diabetic ketoacidosis: type 1 vs type 2 diabetes and the effect of ethnicity. Arch Intern Med. 1999;159(19):2317–22.
42. Jabbar A, Farooqui K, Habib A, Islam N, Haque N, Akhter J. Clinical characteristics and outcomes of diabetic ketoacidosis in Pakistani adults with type 2 diabetes mellitus. Diabet Med. 2004;21(8):920–3.
43. Newton CA, Raskin P. Diabetic ketoacidosis in type 1 and type 2 diabetes mellitus: clinical and biochemical differences. Arch Intern Med. 2004;164(17):1925–31.
44. Liu Z-P, Chen L. Proteome-wide prediction of protein-protein interactions from high-throughput data. Protein Cell. 2012;3(7):508–20.

45. Miryala SK, Anbarasu A, Ramaiah S. Discerning molecular interactions: a comprehensive review on biomolecular interaction databases and network analysis tools. Gene. 2018;642:84–94.
46. Chapelle O, Schölkopf B, Zien A. Semi-supervised learning; 2006.
47. Shin H, Hill NJ, Lisewski AM, Park J-S. Graph sharpening. Expert Syst Appl. 2010;37(12):7870–9.
48. Shin H, Lisewski AM, Lichtarge O. Graph sharpening plus graph integration: a synergy that improves protein functional classification. Bioinformatics. 2007;23(23):3217–24.
49. Tsuda K, Shin H, Schölkopf B. Fast protein classification with multiple networks. Bioinformatics. 2005;21(suppl_2):ii59–65.
50. Wang J. Efficient large margin semisupervised learning. In: Artificial intelligence and statistics; 2007. p. 588–95.
51. Zhu X. Semi-supervised learning literature survey. Comput Sci Univ Wisconsin-Madison. 2006;2(3):4.
52. Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using gaussian fields and harmonic functions. In: ICML; 2003. p. 912–9.

## Publisher's Note