Comorbidity Scoring with Causal Disease Networks

Jong Ho Jhee, Sunjoo Bang, Dong-gi Lee, and Hyunjung Shin

Abstract—In recent years, there has been numerous studies constructing a disease network with diverse sources of data. Many researchers attempted to extend the usage of the disease network by employing machine learning algorithms on various problems such as prediction of comorbidity. The relations between diseases can further be specified into causal relations. When causality is laid on the edges in the network, prediction for comorbid diseases can be more improved. However, not many machine learning algorithms have been developed to concern causality. In this study, we exploit a network based machine learning algorithm that generates comorbidity scores from a causal disease network. In order to find comorbid diseases, semi-supervised scoring for causal networks is proposed. It computes scores of entire nodes in the network when a specific node is labeled. Each score is calculated one at a time and affects to the others along causal edges. The algorithm iterates until it converges. We compared the scoring results of the causal disease network and those of simple association network. As a gold standard, we referenced the values of relative risk from prevalence database, HuDiNe. Scoring by the proposed method provides clearer distinguishability between the top-ranked diseases in the comorbidity list. This is a benefit because it allows the choosing of the most significant ones on an easier fashion. To present typical use of the resulting list, comorbid diseases of Huntington disease and pnuemonia are validated via PubMed literature, respectively.

Index Terms—Causality, comorbidity, disease network, semi-supervised learning

1 Introduction

A disease network consists of disease relations with reference to genome, phenome, metabolome, proteome, and cross-relations among them. Initially and representatively, Goh et al. (2007) proposed a method to build a disease network by considering the fact that similar diseases may share disease related genes [1]. Since then, there have been a number of studies con-structing disease networks [2], [3], [4], [5]. One of the utilities of disease networks is predicting comorbidity. If one contracts a certain disease, it can prevent other diseases that may follow the current one. On the other hand, if we know comorbid diseases, there is a chance to use similar treatment and medication. So far, many comorbidities have been found through biological and clinical researches such as cohort studies or long-term follow-up studies. It is concerned as a rule of thumb. However, but it requires tremendous time and cost. To overcome this difficulty, one can apply machine learning algorithms to predict comorbid diseases given a disease network, which is well curated by the existing studies. Many networks based algorithms can implement a disease network and predict disease comorbidity [6]. To achieve this, sufficient amount of labeled data is required. Technically speaking, labels in a disease network represent whether a patient have certain disease or not. However, most of patients have contracted one or a few diseases. It implies that there are only a few nodes in the network that have labels. Thus, supervised learning algorithms which can be applied to only labeled data, are not suitable for this study. In a circumstance of lack of labeled data, semi-supervised learning (SSL) algorithm would be the adequate method [7], [8]. SSL incorporates not only labeled data but also unlabeled ones to learning. More specifically, graph-based SSL can be the best alternative. Then, when a patient has a certain underlying disease (labeled), it can score the comorbidity for the rest of diseases (unlabeled) in the disease network.

Meanwhile, it is known that causality exists between diseases. For instance, hepatitis can cause liver cancer, but not vice versa [9]. This cause and effect relation is not always reversible generally, and it affects to interactions between diseases in the disease network. So, if possible, it is desirable to add directionality onto the edges in a disease network. Although a great effort on cohort/follow-up studies, causalities between diseases that has been discovered are limited to a very small number. To compensate the deficiency, there are recent studies that attempt to identify causal relationships using a variety of data based on biological, medical, and clinical evidences. Bang et al. (2006) proposed a new approach to define causality between diseases by extracting directional information of genes described in the disease pathways [10]. The study embodies biological mechanisms in identifying disease causalities. Also, Lee and Shin (2017) construct a causal disease network based on text mining was proposed [11]. Unlike other text mining approaches, they used lexicon- and frequency-based analysis to extract causalities from PubMed literature. In fact, cause-and-effect has been significantly concerned in the fields of machine learning and artificial intelligence. In previous studies, Lu and Getoor (2003) suggested link-based algorithm, which

• *The authors are with the Department of Industiral Engineering, Ajou University, 206 World cup-ro, Yeongtong-gu, Suwon, Gyeonggi 16499, South Korea. E-mail: {baical77, smalsunjoo, ldg1226, shin}@ajou.ac.kr.*

Manuscript received 5 Dec. 2017; accepted 3 Mar. 2018. Date of publication 6 Mar. 2018; date of current version 7 Oct. 2019.
(Corresponding author: Jong Ho Jhee).
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2018.2812886

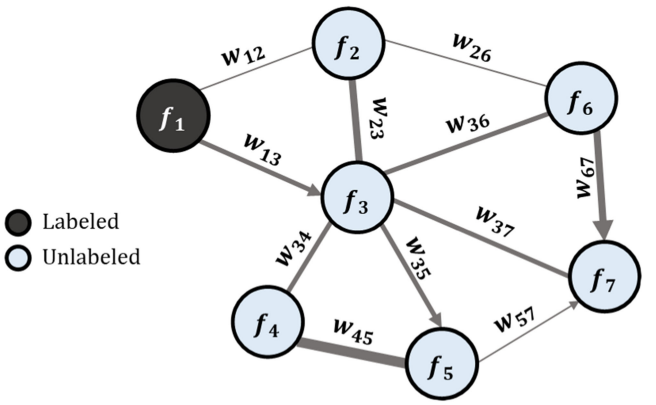


Fig. 1. The causal network structure.

converts the link structure in a causality network into per-node features, and then feed them to logistic regression [12]. In Zhou et al. (2005) [13], they take a hub-authority approach which converts a directed network into an undirected one. They are efficient algorithms if the network is fully directed. However, in a disease network, causalities are scarcely discovered. In such a case, both algorithms treat undirected edges as bi-directed ones, so to speak, one undirected edge is doubly counted as two directed ones. This unnecessarily increases problem complexity. To make matters worse, both algorithms do not preserve the original network structure, which deteriorate readability of the network. In biomedical domain, comprehension on why it happens is never trivial, so readability is essential. To confront the stated difficulties above—lack of labeled data and causality on edges, we propose semi-supervised scoring for causal networks. Hereafter we denote the algorithm as $\mathrm{SSS}^{\mathrm{C}}$. It works on a causal disease network and provides comorbid diseases, without degenerating readability on cause-and-effect between diseases.

The rest of paper is organized as follows. In Section 2, we provide the method to construct a disease network and scoring algorithm, $\mathrm{SSS}^{\mathrm{C}}$. In Section 3, we present experimental comparisons of comorbidity score of association disease networks (ADN) with causal disease networks (CDN). The comorbidity scores by $\mathrm{SSS}^{\mathrm{C}}$ is validated by referencing the values of relative risk (RR) from a well-known disease prevalence database, HuDiNe. We conclude in Section 4.

Semi-supervised scoring for causal graphs

Definition

n: The number of nodes

y_i: Label of node i

μ: Trade off parameter, user specified ($\mu \geq 0$)

w_{ij}: Weight between node i and j

ϵ: User specified threshold for iteration ($\epsilon \geq 0$)

$\boldsymbol{f}_k := \{f_i\}_{i=1}^n$, scores for nodes

Iterate until $\boldsymbol{f}_k$ converges. Start with $\boldsymbol{f}_0 = \mathbf{0}$

do while $(|\boldsymbol{f}_{k-1} - \boldsymbol{f}_k| > \epsilon)$

  for $i = 1{:}n$

$$f_i = \frac{\mu \sum_{i\sim j} w_{ij} f_j + y_i}{\mu \sum_{i\sim j} w_{ij} + 1}$$

  end for

  $k = k + 1$

end while

Fig. 2. Pseudo code of semi-supervised scoring.

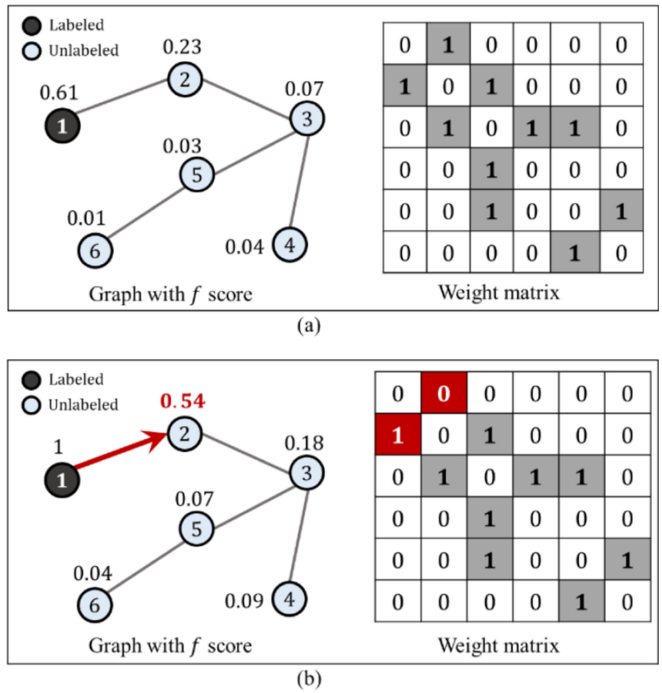


Fig. 3. Toy example of $\mathrm{SSS}^{\mathrm{C}}$ algorithm: (a) Association network and (b) causal network

2 Scoring with Causal Disease Network

The scoring method allows us to find possible co-occurring diseases for a target disease. The way it works is that given a disease network, scores are computed for each labeled

TABLE 1
Data Description

	Number of data	Data sources
Disease	4,663 diseases	MeSH The medical subject headings (www.nlm.nih.gov/mesh/)
Association	78,672 relations between 2,604 diseases 15,777 proteins	PharmDB CTD, GAD, OMIM (http://pharmdb.org/)
Pathway based causality	30,922 genes in 468 pathways resulting 61 causalities between 36 diseases	KEGG Kyoto encyclopedia of genes and genomes (http://www.genome.jp/kegg/pathway.html)
PubMed literature based causality	6,617,833 abstracts resulting 1,011 causalities between 195 disease	PubMed literatures US national library of medicine national institutes of health (www.ncbi.nlm.nih.gov/pubmed)
Prevalence	2,604 prevalence and 266,550 comorbidities of 13,039,018 patients	HuDiNe (http://hudine.neu.edu/)

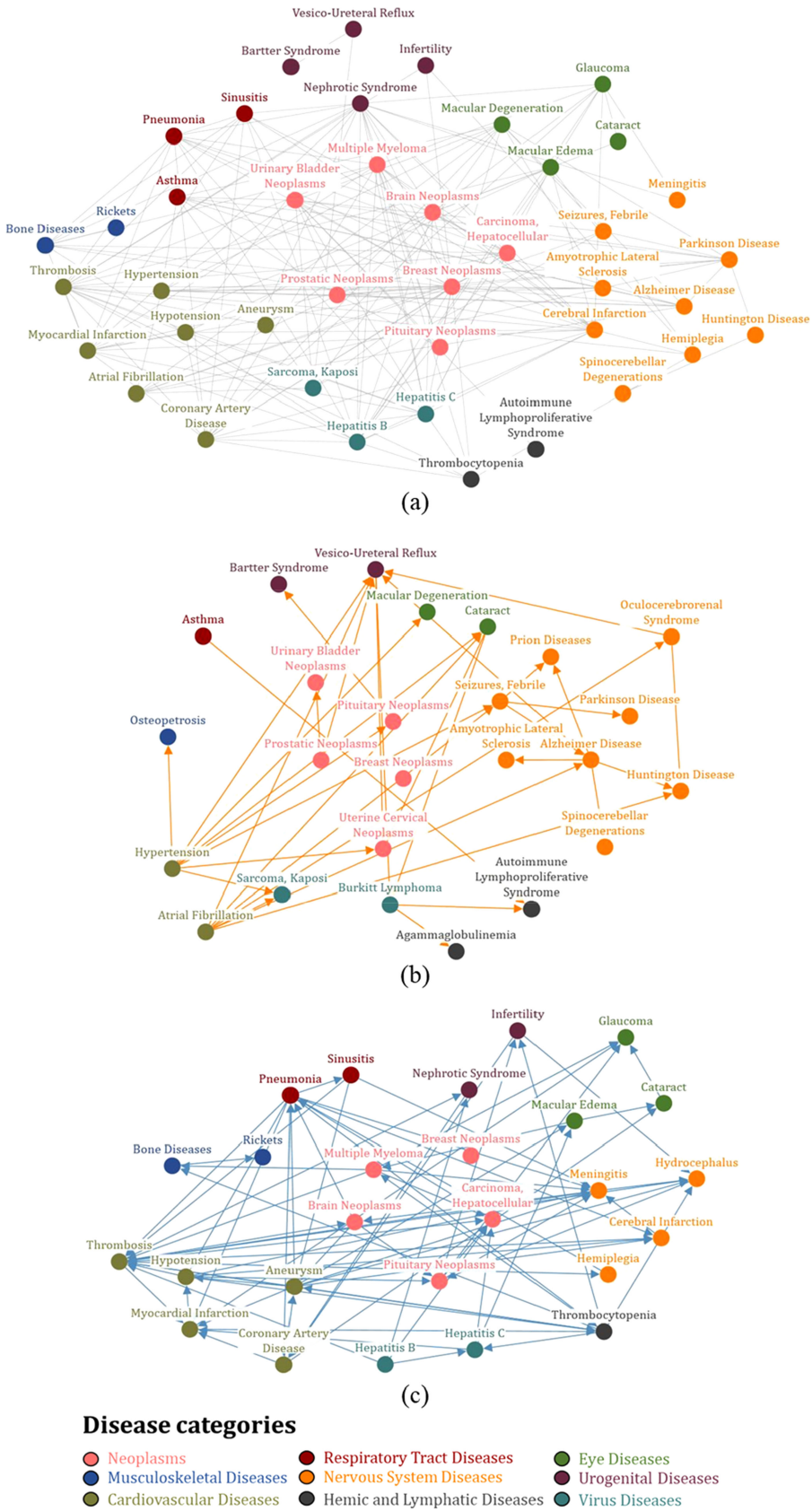


Fig. 4. Disease network: (a) Association disease network (ADN), (b) causality disease network by pathway (CDN^{p}), and (c) causality disease network by literature (CDN^{l}). The colors of nodes represent nine disease categories which are defined by MeSH.

node which is a target disease. Fig. 1 shows the concept of the network structure. 7 nodes can be represented as 7 diseases and labeled node is a target disease. f denotes a score of each disease and it spreads through edges to other diseases in the network. The score increases when connectivity of edges become stronger with associations or causalities.

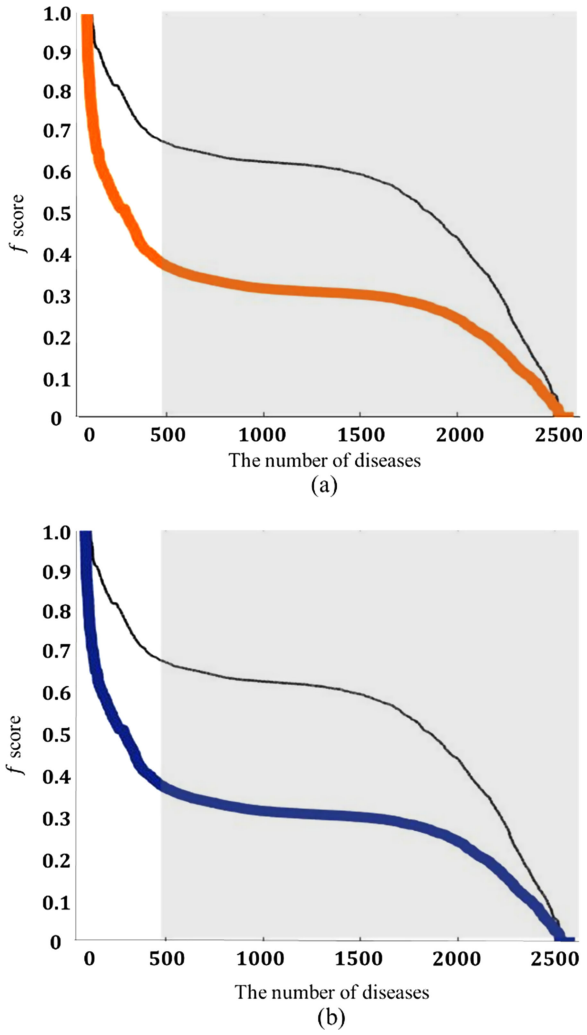


Fig. 5. Comparison of f score of ADN with (a) CDN^p, and with (b) CDN^l, which are sorted in descending order.

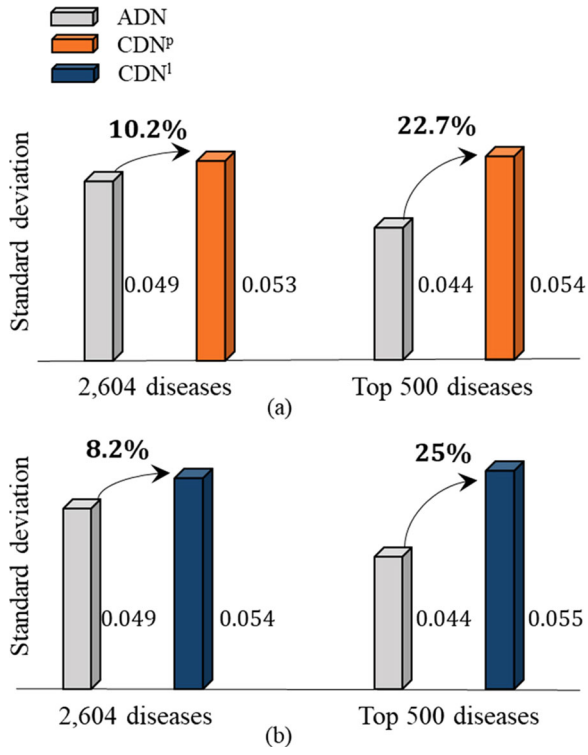


Fig. 6. Standard deviation of f score: (a) CDN^p and (b) CDN^l.

Diseases with high score can be selected as candidates for comorbid diseases. In this study, first, we construct ADN and CDN. After that, we apply SSL and $\mathrm{SSS}^{\mathrm{C}}$ algorithm to ADN and CDN respectively, to compute the scores.

The association disease network is a graph that consists of nodes and edges. Each disease vector $\boldsymbol{X}_{\mathrm{i}} \in \mathbb{R}^{\mathrm{n}}$ is denoted as a node and the association between nodes is denoted as an edge where the similarity between node i and j is represented as w_{ij} of a weight matrix $\boldsymbol{W}$. The disease vectors are based on their shared target proteins. If $\boldsymbol{X}_i$ has a mth protein, then $\boldsymbol{X}_i = \{0, 0, \ldots, X_m = 1, \ldots, 0\}$ Cosine similarity is used to measure similarity be-tween diseases. For the causality disease network data of causalities between diseases are overlaid on ADN. Because of cause and effect relation in causality, the part of edges in CDN are directed. Thus, weight matrix $\boldsymbol{W}$ become asymmetric. More specifically, for node j and i, w_{ij} and w_{ji} have same weight in case of ADN, but $w_{ji} = 0$ and w_{ij} has weight in case of CDN. To overcome this problem, we propose online learning algorithm, $\mathrm{SSS}^{\mathrm{C}}$.

2.1 Scoring Based on Semi-Supervised Learning

Most of biomedical data are unlabeled, because there is massive unknown information. SSL algorithm is optimized for the classification problem of data with these small number of labels by reflecting both labeled and unlabeled data. This is also true of scoring method. It labels a particular node and spreads the score along the edges in the network. Therefore, SSL is suitable for the scoring process. Here, we will apply the algorithm to ADN and CDN.

2.2.1 Association Disease Network

We use graph-based SSL for scoring with ADN. One can obtain f by minimizing the following functional:

$$\min_{f} \mu \boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f} + (\boldsymbol{f} - \boldsymbol{y})^T (\boldsymbol{f} - \boldsymbol{y}), \tag{1}$$

where $\boldsymbol{f} = (f_1, \ldots, f_n)^T$, $\boldsymbol{y} = (0, \ldots, 0, y_i = 1, 0, \ldots, 0)^T$. Only $y_i \in \boldsymbol{y}$ is labeled as 1 for node i among entire nodes. The matrix $\boldsymbol{L}$ is graph Laplacian and μ represents user specified parameter which trades between smoothness and loss. The solution for (1) is:

$$\boldsymbol{f} = (\boldsymbol{I} + \mu \boldsymbol{L})^{-1} \boldsymbol{y}, \tag{2}$$

where $\boldsymbol{I}$ is the identity matrix. SSL can be applied to undirected network such as ADN. The higher score implies the higher similarity to the target disease which can be a high chance of comorbidity between diseases.

2.2.2 Causal Disease Network

It is difficult to calculate comorbidity score in CDN. It has significant fea-ture. Since a similarity matrix of CDN is asymmetric, SSL can not be applied on CDN. To solve this problem, we use online learning SSL algorithm that works in CDN. Unlike SSL, proposing method is not depending on whether similarity matrix is symmetric or asymmetric. More specifically, the weight matrix necessarily changes asymmetrically in order for causality to be reflected. It is

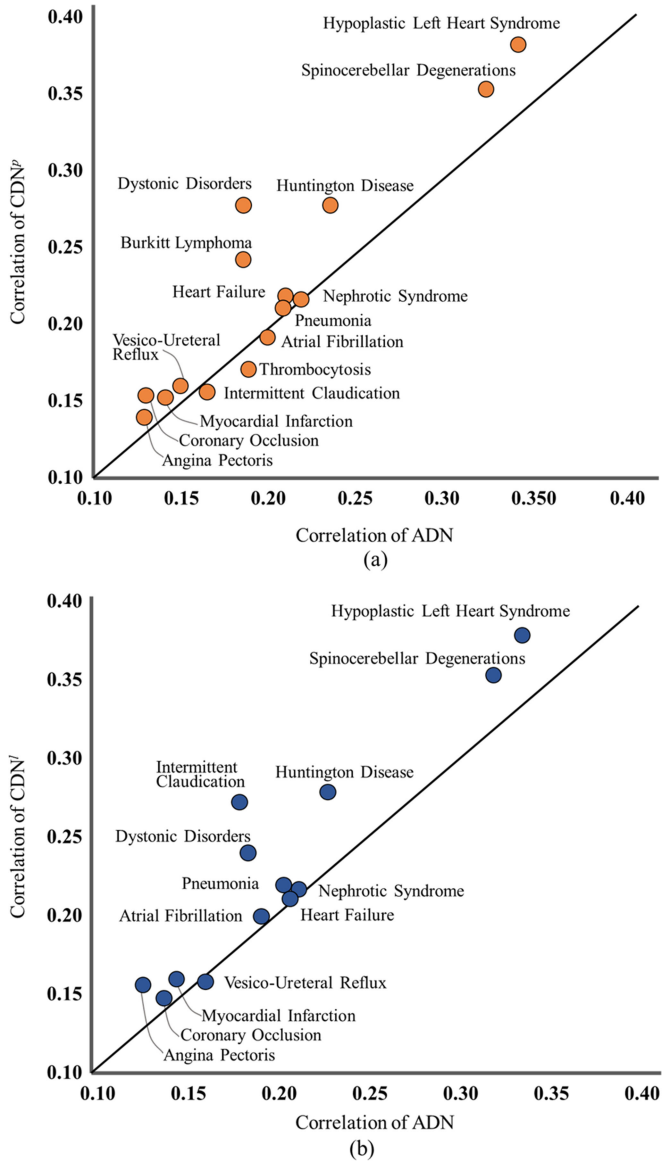


Fig. 7. Correlation between f score and relative risk: (a) CDN^p and (b) CDN^l.

difficult to solve asymmetric matrix with (2), which requires inverse calculation of matrix to derive f scores. Although there are previous studies applying to a causality network, most of them works by converting the causality network into a specific form [12], [13]. $\mathrm{SSS}^{\mathrm{C}}$ algorithm has the advantage of directly calculating the f scores from asymmetric matrix without conversion of CDN. It is flexible to both ADN and CDN.

To obtain f scores in CDN using $\mathrm{SSS}^{\mathrm{C}}$ algorithm, first, we need to minimize the following quadratic functional:

$$\min_f \mu \sum\nolimits_{i\sim j} w_{ij}\left(f_i - f_j\right)^2 + \sum\nolimits_i \left(f_i - y_i\right)^2, \tag{3}$$

where f_i and y_i is an entity of $\boldsymbol{f} = (f_1, \ldots, f_{i-1}, f_i, \ldots, f_{j-1}, f_j, \ldots, f_n)^T$ and $\boldsymbol{y} = (y_1, \ldots, y_{i-1}, y_i, \ldots, y_n)^T$ respectively. w_{ij} represents an element of asymmetric weight matrix $\boldsymbol{W}$. Left term corresponds to the smoothness function S and right term corresponds to the loss function E. Differentiating each term in (3) with respect to f_i, we have a result for minimization in (4):

$$\begin{aligned}
\frac{\delta S}{\delta f_i} &= 2\mu \sum\nolimits_{i\neq j} w_{ij} f_i - 2\mu \sum\nolimits_{i\neq j} w_{ij} f_j \\
\frac{\delta E}{\delta f_i} &= 2\sum\nolimits_I f_i - 2\sum\nolimits_I y_i \\
\frac{\delta(S+E)}{\delta f_i} &= \mu \sum\nolimits_{i\neq j} w_{ij} f_i - \mu \sum\nolimits_{i\neq j} W_{ij} f_i + \sum\nolimits_i f_i - \sum\nolimits_i y_i = 0
\end{aligned} \tag{4}$$

Hence, the solution for the f score is:

$$f_i = \frac{\mu \sum_{i\sim j} w_{ij} f_j + y_i}{\mu \sum_{i\sim j} w_{ij} + 1}, \tag{5}$$

f_i in (5) is computed for each node i in vector $\boldsymbol{f}$ and $i \sim j$ denotes summation for only connected nodes. Second, we need iteration step for the solution. Since we are calculating each f_i of a certain node, it needs to be updated by iteration until $|\boldsymbol{f}_{k-1} - \boldsymbol{f}_k| < \epsilon$ for any $\epsilon \geq 0$. When the iteration ends we can obtain stable f scores. For detail process of the algorithm see the pseudo code in Fig. 2.

Here, we show typical toy example in Fig. 3 to see the effect of causality on f score. Left part of the figure is a network and right part is a weight matrix of corresponding network. There are six nodes in the network and edge weights are simply 1, if there is a relation, or 0, if there is not. Only node 1 is labeled as 1 in this toy example. Because of the causality from node 1 to node 2 in Fig. 3b, node 2 gets higher f score compare to Fig. 3a.

3 Experiments

3.1 Data

To construct disease networks, a list of diseases was collected from the second level of sub category of MeSH of national library of medicine (NLM) [14]. For relations, association and causality information were collected from different sources as shown in Table 1. We showed subset of each disease networks of the collected data in Fig. 4 to improve visibility. In ADN, 2,604 diseases having relations with at least one protein (among 15,777 proteins) were set as nodes with cosine similarity used for calculating weights. (see Fig. 4a) To add causalities to the ADN, results of two previous studies on causal relations between diseases were employed. First, 61 causalities between 36 diseases were obtained from [10], which defines causalities by utilizing metabolic pathways. (see Fig. 4b) Second, 1,011 causalities between 196 diseases were obtained from [11], which quantifies the causal relationship information extracted from the PubMed literatures. (see Fig. 4c) By replacing part of edges in ADN to directed ones (by the causality data above), we were able to construct CDN. We defined CDN by pathway, and literature as CDN^p, CDN^l, respectively. For validation, prevalence data were obtained from HuDiNe that is a database providing clinical records of 13 million patients.

3.2 Enhanced Distinguishability of Scores

The proposed $\mathrm{SSS}^{\mathrm{C}}$ have been applied to the three disease networks, ADN, CDN^p, and CDN^l, to calculate f score for the rest of 2,603 diseases when a certain disease has been set as label '1'. Figs. 5a and 5b show results for experiments to compare f score of ADN (black), for a certain disease, with

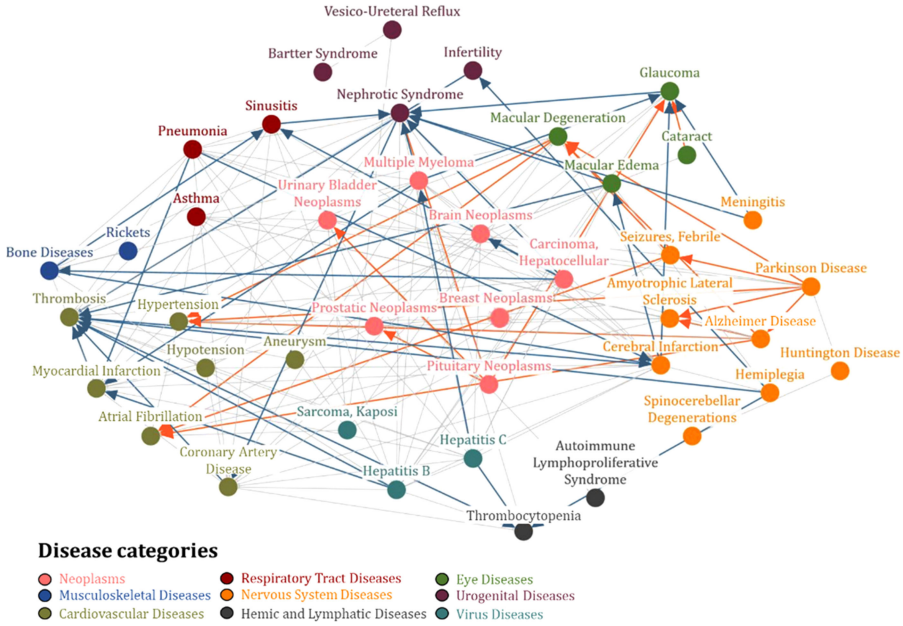


Fig. 8. Causality network for Figs. 4b and 4c overlaid on the Fig. 4a, which means the causality is reflected on the original network.

CDN^p (orange) and with CDN^l (blue) respectively. These two results show that when the causality is added to the network the standard deviation of the score distribution is relatively larger even though they may have the same priority for comorbidity. That is, the distinguishability for priority improves by increased discernment between scores. Moreover, the diseases with high score are our interest and relatively the subordinate diseases with low score are not our concern. In both Figs. 5a and 5b, from disease 1 to 500 the graphs with causality have sharper slope than that of ADN. As shown in Fig. 6, the standard deviation of total 2,604 diseases hardly has difference between ADN (0.049) and CDN^p (0.053). Rate of difference is 10.2 percent. However, there is significant difference in top 500 disease. Standard deviation is 0.044 for ADN and 0.054 for CDN^p, and rate of difference is 22.7 percent (see Fig. 6a). In the same vain, rate of difference is 8.2 percent and 25 percent CDN^l. (see Fig. 6b). As a result, the distinguishability of scores for predicting comorbidity are enhanced by adding causality between diseases and by focusing on the top scored diseases.

3.3 Validation with Prevalence Data

There are clinical records which is known as gold standard of comorbidity between diseases. HuDiNe collected prevalence and concurrent occurrence information, so-called comorbidity from patients' records, of approximately 13 million of patients. We need to evaluate whether the score of comorbidity from the proposed method follows the gold standard or not. We used relative risk which is the probability of developing other diseases associated with a certain disease. Then, RR was compared with f score of three disease network for sampled 15 target diseases ($\mathrm{correlation} > 0.12$, $\mathrm{p\text{-}value} < 0.005$). Fig. 7a shows the result of comparing rank correlation between RR and f score of ADN with correlation between RR and f score of CDN^p. In the same manner, Fig. 7b indicates the results of CDN^l. Obviously, f score of CDN has stronger correlation with RR.

As a validation, we searched medical literatures related to the high scored comorbidity, such as Huntington disease and pneumonia.

Huntington disease and landau-kleffner syndrome (LKS): There is a report that neurological disorder including Huntington disease and mitochondrial disorders are associated with childhood dementia such as LKS [15]. A child with LKS have the loss of verbal expression and language comprehension.

Huntington disease olivopontocerebellar atrophies (OPCA): Also, OPCA is neurodegenerative disease manifesting mainly cerebellar and brainstem symptoms [16]. Thus, OPCA is considerably related to brain atrophy found in Huntington disease [17].

Pneumonia and bacterial meningitis: There is a report showing community-acquired bacterial meningitis has a high rate of an unfavorable outcome in adults (34 percent). Finally, factors predictive of pneumococcal infection were associated with an unfavorable outcome (advanced age; presence of otitis or sinusitis, pneumonia, or immunocompromised status; and absence of rash) [18].

Pneumonia and leukocyte adhesion deficiency: In this report, the clinical features of the type 2 leukocyte adhesion deficiency syndrome have recently been described. The two cases report here involved two unrelated boys three and five years old, each the offspring of consanguineous parents. Both have severe mental retardation, short stature, a distinctive facial appearance, and the Bombay (hh) blood phenotype, and both have had recurrent episodes of bacterial infection, mainly pneumonia, periodontitis, otitis media, and localized cellulitis without the formation of pus [19].

Fig. 8 shows disease network with association overlaid by both pathway based causality and PubMed based causality. The comorbidity scores are obtained from this network and it is sorted in descending order as shown in Fig. 9. Red colored circle represents the target disease and grey colored circle represents the gold standard. Such as refsum disease or friedreich ataxia in (a) and urinary bladder calculi or pericarditis in

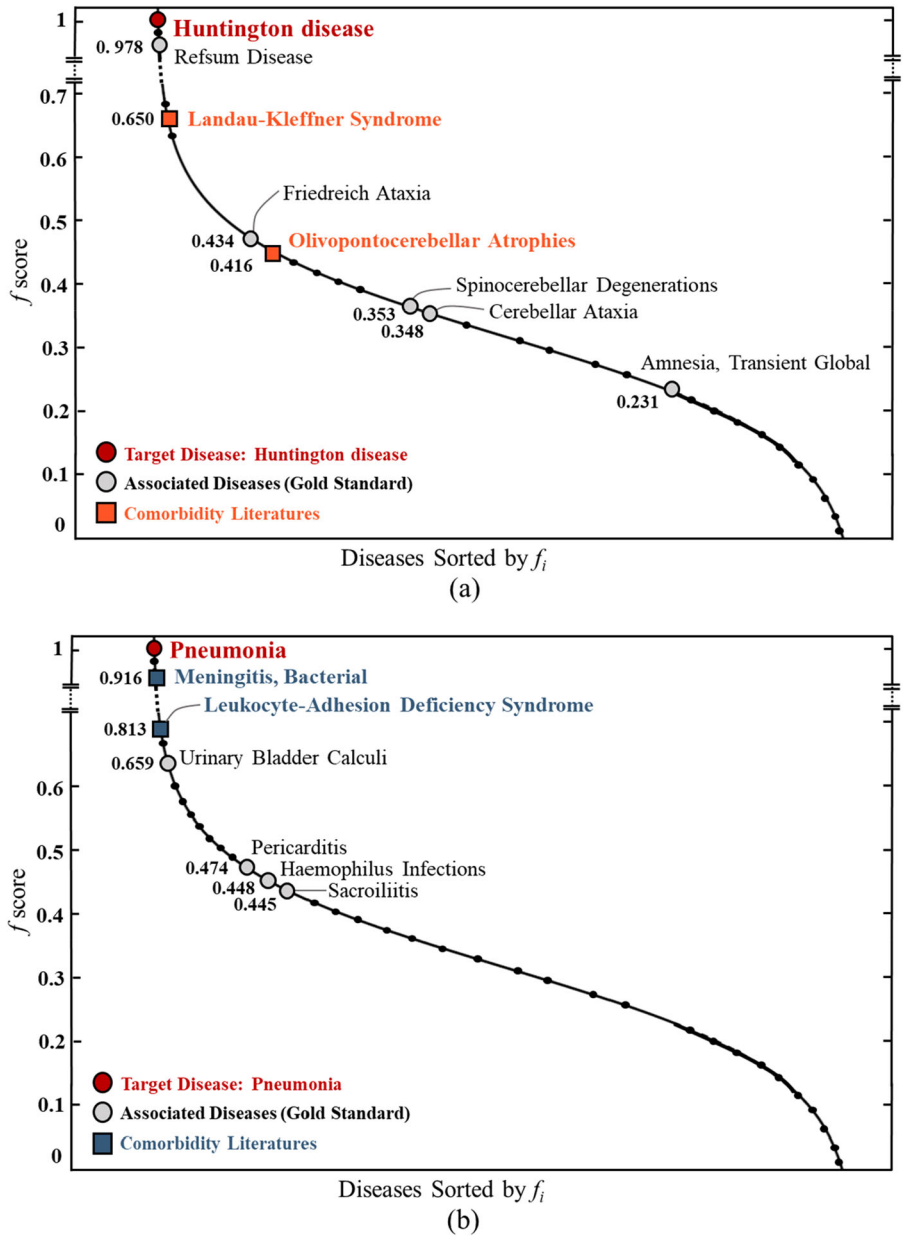


Fig. 9. Comorbidity of target diseases: (a) Huntington disease and (b) pneumonia.

(b) were matched with gold standard. Orange colored squares in (a) are the ones we validated with the literatures for Huntington disease and blue colored squares in (b) are the ones we validated with the literatures for pneumonia.

4 Conclusion

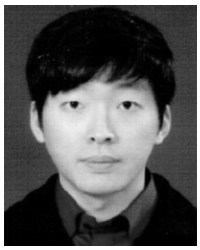
The research proposes a method to score comorbidity with causal disease network. semi-supervised scoring for causal networks (SSS^c) algorithm can reflect directional information naturally without conversion of the network. SSL and SSS^c are used to obtain comorbidity scores from ADN and CDN. In the experiment, disease scores are calculated for each target disease. Standard deviation of f scores is compared between ADN and CDN. For top 500 disease, standard deviation of CDN is much larger than ADN. As a result, it is shown that scoring comorbidity by reflecting even a few known causalities enhances distinguishability. We have validated the resulting score of CDN is reasonable by comparing with the relative risk. Among sampled 15 diseases that have strong correlation with relative risk, Huntington disease was selected as a target disease. 3 high scored diseases are chosen and verified with literatures. This research has novelty in following aspects. The method is proposed to overcome unnecessarily increasing problem and to preserve the original network structure, which can sustain readability of the network. Not only examples we mentioned but also other cases can be found by comparing the comorbidity scores. In further researches, we are planning to improve computation time of SSS^C algorithm. It is difficult to verify causal relationships as well as disease associations. Although the results of this study were not all verifiable, some of the known relationships were verified using a clinical record and PubMed literatures in a circumstantial way. It is expected that the results of this study will be used as a subsidiary for the qualitative research of the clinician and will verify more relationships in the future.

ACKNOWLEDGMENTS

The authors would like to gratefully acknowledge support from the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2015R1D1A1A01057178), ICT R&D program of MSIP/IITP (No. 2017-0-00887) and the Ajou University research fund. J. H. Jhee and S. Bang are the joint first authors.

REFERENCES

[1] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proc. Nat. Academy Sci. USA*, vol. 104, no. 21, pp. 8685–8690, 2007.
[2] C. A. Hidalgo, et al., "A dynamic network approach for the study of human phenotypes," PLoS computational biology 5.4 (2009): e1000353.
[3] D. A. Davis and N. V. Chawla, "Exploring and exploiting disease interactions from multi-relational gene and phenotype networks," *PloS One*, vol. 6, no. 7, 2011, Art. no. e22670.
[4] Y. Li and P. Agarwal, "A pathway-based view of human diseases and disease relationships," *PloS One*, vol. 4, no. 2, 2009, Art. no. e4346.
[5] X. Zhang, R. Zhang, Y. Jiang, P. Sun, G. Tang, X. Wang, H. Lv, and X. Li, "The expanded human disease network combining protein–protein interaction information," *Eur. J. Human Genetics*, vol. 19, no. 7, pp. 783–788, 2011.
[6] F. He, G. Zhu, Y.-Y. Wang, X.-M. Zhao, and D.-S. Huang, "PCID: A novel approach for predicting disease comorbidity by integrating multi-scale data," *IEEE/ACM Trans. Comput. Biology Bioinf.*, vol. 14, no. 3, pp. 678–686, May/Jun. 2017.
[7] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. Cambridge, USA: MIT Press, 2006.
[8] J. Kim and H. Shin, "Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data," *J. Amer. Med. Infor. Assoc.*, vol. 20, no. 4, pp. 613–618, 2013.
[9] P. P. Anthony, "Precursor lesions for liver cancer in humans," *Cancer Res.*, vol. 36, no. 7, Part 2, pp. 2579–2583, 1976.
[10] S. Bang, J.-H. Kim, and H. Shin, "Causality modeling for directed disease network," *Bioinformatics*, vol. 32, no. 17, pp. i437–i444, 2016.
[11] D.-g. Lee and H. Shin, "Disease causality extraction based on lexical semantics and document-clause frequency from biomedical literature," *BMC Med. Inf. Decision Making*, vol. 17, no. 1, 2017, Art. no. 53.
[12] Q. Lu and L. Getoor, "Link-based classification," *Proc. 20th Int. Conf. Mach. Learn. (ICML)*, 2003, pp. 496–503.
[13] D. Zhou, T. Hofmann, and B. Schölkopf, "Semi-supervised learning on directed graphs," *Adv. Neural Inform. Process. Syst.*, 2004 pp. 1633–1640.
[14] H. J. Lowe, and G. O. Barnett, "Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches," *Jama*, vol. 271, no. 14, pp. 1103–1108, 1994.
[15] C. E. Coffey and R. A. Brumback, *Pediatric Neuropsychiatry*. Baltimore, MD, USA Lippincott Williams & Wilkins, 2006.
[16] V. N. Kornienko and I. N. Pronin, "Diagnostic neuroradiology," *Am. Soc. Neuroradiology*, vol. 30, no. 9, 2009, Art. no. E136.
[17] G. Halliday, D. McRitchie, V. Macdonald, K. Double, R. Trent, and E. McCusker, "Regional specificity of brain atrophy in Huntington's disease," *Exp. Neurology*, vol. 154, no. 2, pp. 663–672, 1998.
[18] D. Van de Beek, J. de Gans, L. Spanjaard, M. Weisfelt, J. B. Reitsma, and M. Vermeulen, "Clinical features and prognostic factors in adults with bacterial meningitis," *New England J. Med.*, vol. 351, no. 18, pp. 1849–1859, 2004.
[19] A. Etzioni, M. Frydman, S. Pollack, I. Avidor, M. L. Phillips, J. C. Paulson, and R. Gershoni-Baruch, "Recurrent severe infections caused by a novel leukocyte adhesion deficiency," *New England J. Med.*, vol. 327, no. 25, pp. 1789–1792, 1992.
[20] Home - MeSH - NCBI, National Center for Biotechnology Information, U.S. National Library of Medicine, http://www.ncbi.nlm.nih.gov/mesh
[21] Pharm DB-K. [Online]. Available: http://www.pharmdb.org.
[22] "HuDiNe. Human disease network," Accessed: Nov 1, 2014. [Online]. Available: http://www.alphaminers.net/.



Jong Ho Jhee received the BS degree from Yonsei University in 2013, and is currently pursuing the PhD degree in the Graduate School of Industrial Engineering, Ajou University, South Korea. His current research interest is on machine learning and especially causal inference in biomedical informatics.

Sunjoo Bang received the BA degree from Ajou University in 2014, and is currently pursuing the PhD degree in the Graduate School of Industrial Engineering, Ajou University, South Korea. Her research interests are in network algorithms and machine learning algorithms for biomedical domain issues; disease causality definition, or target gene identification, Alzheimer's disease diagnostic model, and so on.

Dong-gi Lee received the MS degree from Ajou University in 2016, and is currently pursuing the PhD degree in the Graduate School of Industrial Engineering, Ajou University, South Korea. His current research interest is on biomedical informatics and historical inference using various techniques of machine learning algorithms.

Hyunjung (Helen) Shin received the PhD degree in data mining from Seoul National University, and further majored in machine learning during her Post-Doc at the Max Planck Institute in Germany. Since 2006, she has joined Ajou University as a faculty member of the Department of Industrial Engineering. Her theory interest is more focused on data mining algorithms including machine learning. Her research activities on application range across areas as different as biomedical informatics, hospital fraud detection, direct marketing in CRM, Oil/Stock price prediction, etc.