Bioinformatics, 32, 2016, i437–i444
doi: 10.1093/bioinformatics/btw439
ECCB 2016

OXFORD

Causality modeling for directed disease network

Sunjoo Bang†, Jae-Hoon Kim† and Hyunjung Shin*

Department of Industrial Engineering, Ajou University, Wonchun-Dong, Yeongtong-Gu, Suwon 443-749, South Korea

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Abstract

Motivation: Causality between two diseases is valuable information as subsidiary information for medicine which is intended for prevention, diagnostics and treatment. Conventional cohort-centric researches are able to obtain very objective results, however, they demands costly experimental expense and long period of time. Recently, data source to clarify causality has been diversified: available information includes gene, protein, metabolic pathway and clinical information. By taking full advantage of those pieces of diverse information, we may extract causalities between diseases, alternatively to cohort-centric researches.
Method: In this article, we propose a new approach to define causality between diseases. In order to find causality, three different networks were constructed step by step. Each step has different data sources and different analytical methods, and the prior step sifts causality information to the next step. In the first step, a network defines association between diseases by utilizing disease–gene relations. And then, potential causalities of disease pairs are defined as a network by using prevalence and comorbidity information from clinical results. Finally, disease causalities are confirmed by a network defined from metabolic pathways.
Results: The proposed method is applied to data which is collected from database such as MeSH, OMIM, HuDiNe, KEGG and PubMed. The experimental results indicated that disease causality that we found is 19 times higher than that of random guessing. The resulting pairs of causal-effected diseases are validated on medical literatures.
Availability and Implementation: http://www.alphaminers.net
Contact: shin@ajou.ac.kr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Causality between diseases is an important concept when practicing patients in medicine. For example, causality that insulin resistance causes type 2 diabetes, which is verified by cohort study (Lillioja *et al.*, 1993) can be realized in practice to treat a patient suffering from resistance of insulin. In order to prevent him or her from being deteriorated to diabetes mellitus type 2, medical experts should make efforts to adjust level of insulin in the blood (Saltiel and Olefsky, 1996). Mostly causalities between diseases, however, are verified by cohort studies (Hemingway and Marmot, 1999; Kukull *et al.*, 2002; Lillioja et al., 1993; McDonald *et al.*, 1993). While cohort studies can get results with high objectivity, they require high cost and time consumption. Therefore, to overcome such weaknesses, it is beneficial to define causality between diseases based on diversified bio-medical data. These types of data are well-curated by experts and relatively easy to process information. From many cases in conventional studies which define relations between diseases using genetics and molecular biological data, Goh *et al.* (2007) proposed a method to build disease network based on relations between disease and disease-related gene. And also there are methods which expressed association between diseases as network using genetic character, phenotype, protein interaction and metabolic pathway (Davis and Chawla, 2011; Hidalgo *et al.*, 2009; Li and Agarwal, 2009; Zhang *et al.*, 2011). However, existing researches are only limited to clarify *association* of two diseases but *causality*. Li and Agarwal (2009) proposed an association network between diseases which takes into account metabolic pathway. In addition, there are some studies which developed systematic methods to define metabolic pathway (Schuster *et al.*, 1999, 2000). However, most of such studies are not interested in reflecting directional information between genes in the pathways which are interactive in cells or focused on complicated methods to understand metabolic pathway only. By some roundabout way, study of Xu *et al.* (2014) took advantage of text mining method to extract causality between diseases in medical literatures.

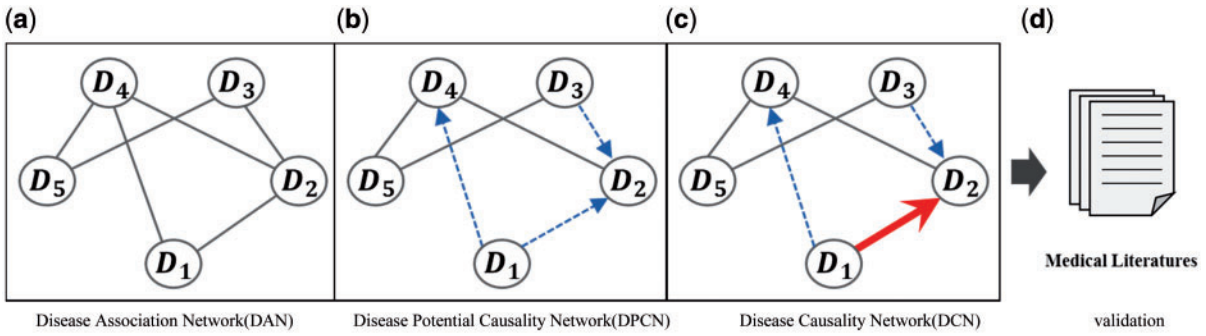

Fig. 1. *Proposed method*: in this article, we propose a new approach to define causality between diseases. In order to find causality, three different networks were constructed step by step. Each step has different data sources and different analytical methods, and the prior step sifts causality information to the next step. (a) A disease association network (DAN), (b) a disease potential causality network (DPCN), (c) a disease causality network (DCN) and (d) a validation on related medical literatures

In this study, we propose a novel method to define causality between two diseases from gene, clinical and metabolic pathway information. Most notably, we provide a method of systematically analyzing metabolic pathways to extract causality. Comparing with the existing study, we expect that causality found from molecular and biological data may be more fundamental and reliable. Using those diverse sources and types of molecular and clinical information, we define causal relation between diseases step by step which is consists of *Association*, *Potential Causality* and *Causality*. Each step provides results as a form of network, and the resulting network of the prior step reduces spurious information for the next one. The proposed method is described in Figure 1. First, Disease Association is defined as shown in Figure 1a. Each node indicates diseases whereas edge to connect two nodes indicates simple association. Association between diseases is the most important precondition which need to be taken into account in order to define causality as Bradford hill argues (Hill, 1965). In the following step, Disease Potential Causality is endowed to, as shown in Figure 1b, those diseases whose association is defined in the previous network. Prevalence and comorbidity information is used in this step: the concept of relative risk between diseases can be calculated by adopting the existing study (McNutt *et al.*, 2003; Zhang and Kai, 1998). However, relative risk is only a scale to measure strength of association, but cannot be used to represent causality by itself. Therefore, we propose a formula calculating potential causality between diseases. We use a term 'potential' to avoid confusion that the causality found from clinical information is not yet confirmed by information from biomolecular level. Finally, information extracted from disease-related metabolic pathway confirms disease causality. Among the pairs of diseases whose edge having potential causality, most relevant ones are selected based on pathway information. Figure 1b and c exemplifies this process: D_1 has potential causality with D_2 and D_4 in the previous step, but only causality with D_2 is supported by pathway information. And therefore, it remains as a confirmed causality.

The proposed method is applied to data which is collected from database such as MeSH, OMIM, HuDiNe and KEGG. And the pairs of cause-effected diseases are validated though searching the biomedical reports from PubMed.

2 Methods

This study proposes a method to construct disease network to verify causality. Each network which consists of three steps uses a different data and analysis method depending on their purposes. The first step, *Disease Association Network (DAN)* defines association from disease-gene related data while *Disease Potential Causality Network (DPCN)* defines potential causality using prevalence and comorbidity information. Finally, *Disease Causality Network (DCN)* confirms causality between diseases through gene directional information extracted from metabolic pathway.

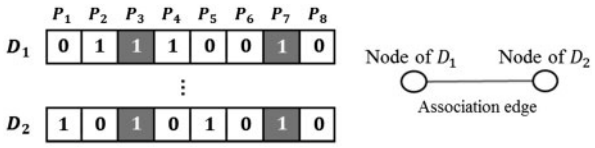

Fig. 2. Disease association network (DAN) construction method

2.1 Disease association network construction

DAN has a similar structure as that of The Human Disease Network (HDN) proposed by Goh *et al.* (2007). To construct DAN, we compare disease–protein vectors of two diseases: if two diseases share at least one protein, the edge between the two diseases is connected and represents their association. The strength of association is assigned proportional to the number of proteins which are shared by the two diseases. Figure 2 describes disease–protein vectors of D_1 and D_2, and association strength (AS) is calculated as 2 because they share two proteins. Likewise, the whole network is constructed.

2.2 Disease potential causality network construction

In order to form DPCN, we use prevalence and comorbidity information indicating frequency of each disease occurrence and concurrent occurrence of two diseases respectively. Figure 3 shows the proposed approach to applying relative risk to define potential causality from the prevalence and comorbidity.

Relative risk (RR) is defined as Equation (1) assuming that risk factor is $'A'$ while an incident is $'B'$ (McNutt et al., 2003; Zhang and Kai, 1998). If RR is larger than 1, then $'A'$ becomes cause of $'B'$ and $'B'$ is effect of $'A'$.

$$\mathrm{RR}(A,B) = \frac{p(B|A)}{p(B|\sim A)} \tag{1}$$

Given two diseases, two values of RR exist: one can be causal disease to the other, and vice versa. Therefore, we compare the two values and selects larger one as a potential causality values between two diseases. To sophisticate this, we reflect of ratio of RR to the formula in order to reduce sensitivity depending on difference of

two RR values. It is defined in Equation (2), and denoted as potential causality strength (PCS).

$$\mathrm{PCS}(D_1, D_2) = \varphi(\mathrm{RR}(D_1, D_2) - \mathrm{RR}(D_2, D_1)) \cdot \mathrm{RRR}(D_1, D_2) \quad (2)$$

where $\varphi(u) = \left\{ \begin{array}{c} 1 \ if \ u > 0 \\ 0 \ otherwise \end{array} \right\}$

and $\mathrm{RRR}(D_1, D_2) = \frac{\mathrm{RR}(D_1, D_2)}{\mathrm{RR}(D_2, D_1)}$.

Figure 3 shows an example of the calculation of Equation (2). Assume that prevalence of disease D_1 and D_2 is 45 and 20, respectively and comorbidity that two diseases occurs simultaneously is 15. Under reserved condition, $\mathrm{RR}(D_1, D_2)$ is 3.6 and $\mathrm{RR}(D_2, D_1)$ is 2 according to Equation (1). In this case, because the former is larger than the latter, D_1 becomes a causal disease while D_2 is an effected disease. The potential causal strength between them is 1.8 by Equation (2). This means that the risk of 'D_1 may cause D_2' is approximately 1.8 times higher than that of the opposite case.

2.3 Disease causality network construction

DPCN defined by previous step verifies existence of causality for disease. Based on this, DCN is formed from metabolic pathway data: Figure 4 indicates processes to form DCN. It consists of the sequence of (a) sharing block, (b) flow function and (c) causality function.

2.3.1 Sharing block

A disease is thought of a result of gene mutations that cause disruptions in underlying cellular functions. If two diseases are related, the respective disease-pathways are highly likely to be related. In a word, a common set of genes exists between two disease pathways. This can incur chain reactions: one gene belonging to a pathway is disrupted, it will lead to disruption of the common gene set, and eventually will disrupt the genes even solely belonging to the pathway of other disease. In order to reflect this idea, we first find the commonly existing genes (sharing genes) in the two metabolic pathways and set them as a block. Between diseases, directional information between shared genes is ignored when causality is defined. However, the shared genes in the block become criteria to calculate causal influence between two diseases. To gather the directional information to the block from unique genes—genes belonging to only one disease, it requires a pointing reference, and the block of the shared genes play such a role. Figure 4a shows two metabolic pathways of disease D_1 (left) and disease D_2 (right), and presents the block of shared genes and the directionalities to it from either D_1 or D_2 in terms of disease unique genes.

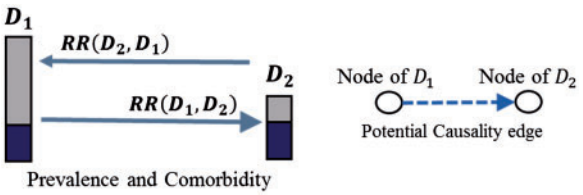

Fig. 3. Disease potential causality network (DPCN) construction method

2.3.2 Flow function

As genes in a metabolic pathway have complicated and entangled directional aspects. To boil down the complexity of directionality in the pathway, we define a flow function as in Equation (3). By calculating the value of Equation (3) one can figure out whether the disease unique genes are influencing or being influenced by the sharing block. A positive value stands for influencing direction to the block, whereas a negative value means influenced direction from the block. For this, the flow function is decomposed of inflow function and outflow function:

$$\mathrm{Flow}(D_1|D_2) = \sum_i^{n_I} \mathrm{IN}_i(D_1|D_2) - \sum_k^{n_O} \mathrm{OUT}_k(D_1|D_2) \quad (3)$$

where $'D_i|D_j'$ stands for 'Inflow/outflow' of $D_i$toward/from the sharing block of D_i and D_j. And, n_I and n_O defines the multiple paths of genes to the block. Inflow and outflow function is defined as follows:

$$\mathrm{IN}(D_1|D_2) \ = \sum_{l=1}^{L_I} S(l), \ \ \mathrm{OUT}(D_1|D_2) = \sum_{l=1}^{L_O} S(l) \quad (4)$$

where $S(l) = \mathrm{sign}(l)\exp^{-(l-1)}$ and $\mathrm{sign}(l) \in \{+1, -1\}$.

The respective function uses exponential function $S(l)$ so that influence from/to the sharing block is diminished as a gene is larger away from it.

L_I and L_O define the length paths of genes from/to the block. Figure 4b describes the above process.

2.3.3 Causality function

Given two values of flow function between two diseases, $\mathrm{Flow}(D_1|D_2)$ and $\mathrm{Flow}(D_2|D_1)$, disease causality can be calculated by comparing them. Equation (5) shows the causality function.

$$\mathrm{Causality}\ (D_1, D_2) = \left\{ \begin{array}{c} 1 \text{ if } \mathrm{Flow}(D_1|D_2) > \mathrm{Flow}(D_2|D_1) \\ -1 \text{ else if } \mathrm{Flow}(D_1|D_2) < \mathrm{Flow}(D_2|D_1) \\ 0 \text{ otherwise} \end{array} \right. \quad (5)$$

If the value of $\mathrm{Causality}(D_1, D_2)$ is '+1', it stands for D_1 is causal disease of D_2. If the value is '−1', D_1 becomes the effected

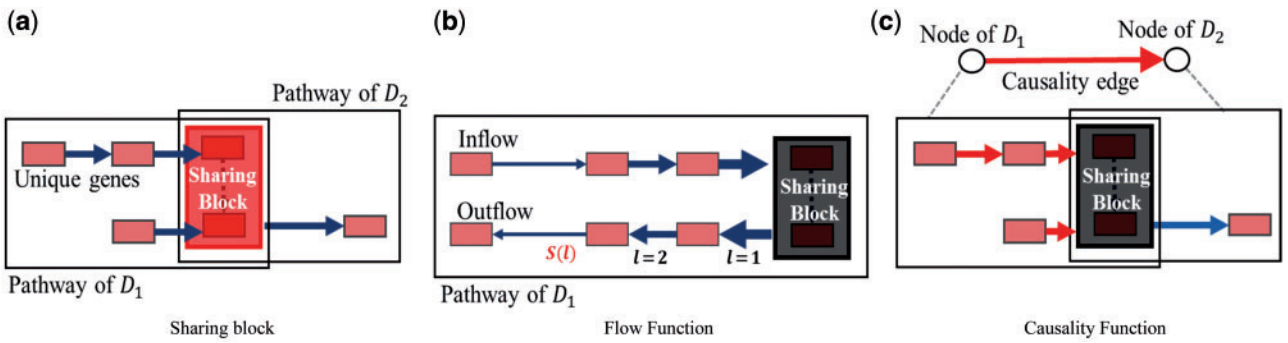

Fig. 4. Disease causality network (DCN) construction method

disease being influenced from D_2. Causality strength is defined as follows.

$$CS(D_1, D_2) = \max \{ \text{Flow}(D_1|D_2), \text{Flow}(D_2|D_1) \} \quad (6)$$

A toy example for the process of extracting causality between insulin resistance and type 2 diabetes mellitus from metabolic pathways is described below (Fig. 5).

Five genes, INS, INSR, P13K, IRS and SOCS, which are common in both disease constitutes are in the sharing block. IRS1, GLUT4 and mTOR are unique disease genes of type 2 diabetes mellitus, Whereas LAR, PTP1B, JNK and IKK are those of insulin resistance. The values of flow function, Flow(IR| T2DM), is 5, and Flow(T2DM| IR) is −3 according to Equation (3). Therefore, causality value of the disease pair, Causality(IR, T2DM) becomes 1 according to Equation (5), which implies that 'insulin resistance causes type 2 diabetes mellitus', and the corresponding strength of Equation (6) is 5. In practice, however, a disease can be entangled with several pathways, which is more complex than the above toy example. Then, the mechanism in Equation (5) can be similarly applied to all possible pairs of combinations of the pathways.

3 Experiments

3.1 Data

In order to verify proposed method, we used data summarized Table 1. Disease nodes belonging to the proposed network is collected from MeSH that is Thesaurus database for medical areas specified by United States National Library of Medicine. We collected 4663 diseases which are defined by the second level of sub category of MeSH diseases.

As shown in the previous sections, we constructed three different networks depending on the data sources we used. Accordingly, edges in each network are different. Disease–protein information which forms edge of DAN is acquired from PharmDB. PharmDB defines relations between various database such as the Comparative Toxicogenomics Database(CTD), Entrez Gene Interactions and Online Mendelian Inheritance in Man(OMIM). From 15 777 dimensional binary vectors of disease–protein information (protein that is related to disease has '1' otherwise it has '0'), 2604 diseases are connected in DAN. In order to form edges of DPCN, we collected prevalence and comorbidity information which are provided by HuDiNe. HuDiNe is database which processed records of approximately 13 million of patients and this consists of 2604 prevalence and 266 550 comorbidity between two diseases for 2604 MeSH diseases. For DCN, we acquired disease–pathway relation information from KEGG. KEGG is a database which provides association of molecule and reaction processes for diseases via pathway map that is manually drawn. About 207 pathways are collected, which are related to 163 diseases.

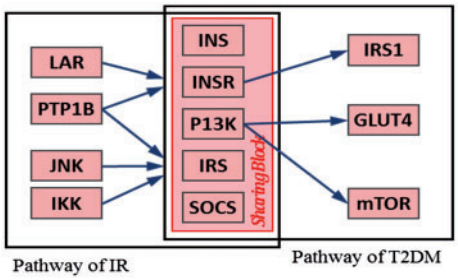

Fig. 5. A toy example for the process of extracting causality

3.2 Result of construction of disease networks

Table 2 indicates results of summarizing relations between DAN, DPCN and DCN. Since each network is structured step by step, DPCN is structured for a pair of disease whose association is defined through DAN, and DCN is formed for a pair of diseases whose potential causality is clarified through DPCN and then final causality is confirmed.

As the result, DAN defined 738 402 association between 2604 diseases and extracted 133 261 potential causalities between 1015 diseases. And, finally this extracted 61causalities between 36 diseases from DCN.

Figure 6 shows DAN and DPCN's sub-network for 36 diseases whose causality is defined in DCN. Each node size is proportional to number of other connected nodes while edge is proportional to strength of association or causality. The different colors of nodes identify disease categories according to eight classifications of diseases defined by MeSH.

Table 2. Construction result of each network

	DAN	DPCN	DCN
Number of node	2604	1015	36
Number of edge	738 402	133 261	61
Network density	22%	0.4%	0.000018%

Table 1. Data for diseases, disease–protein relationship, prevalence, comorbidity, metabolic pathway and literature

	Disease association network	Disease potential causality network	Disease causality network	Validation
Node	Disease MeSH The Medical Subject Headings (www.nlm.nih.gov/mesh/) 4,663 diseases			PubMed literature US National Library of Medicine National Institutes of Health, (www.ncbi.nlm.nih.gov/pubmed)
Edge	Disease–protein relationship PharmDB CTD, GAD, OMIM (http://pharmdb.org/) 78 672 relations between 2604 diseases and 15 777 proteins	Prevalence and comorbidity HuDiNe (http://hudine.neu.edu/) 2604 prevalence and 266 550 comorbidities of 13 039 018 patients	Metabolic pathway KEGG (http://www.genome.jp/kegg/pathway.html) 30 922 genes among 468 pathways	

The number in parentheses indicates the amount of data originating from the respective sources.

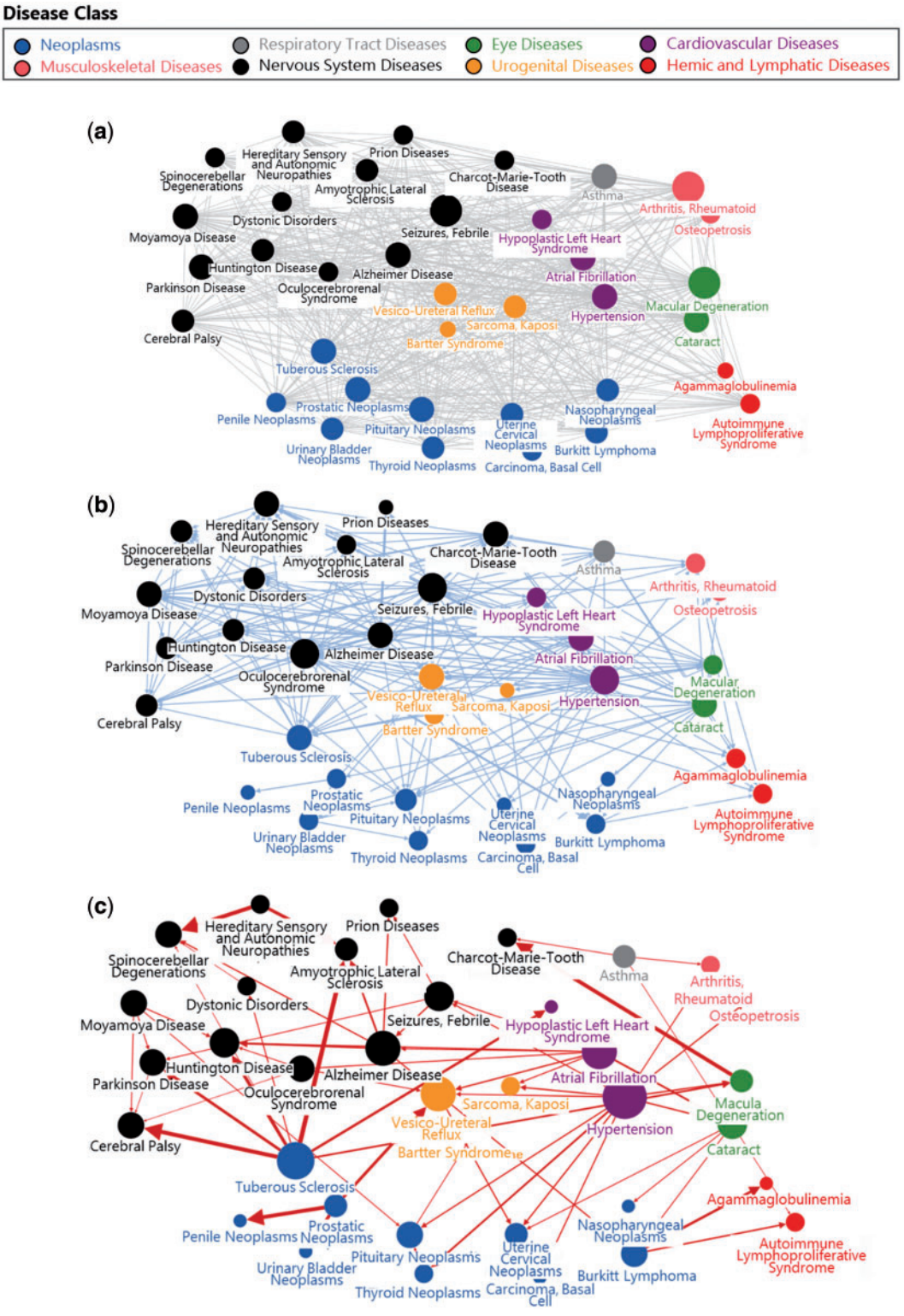

Fig. 6. *Construction of proposed stepwise disease network with 36 diseases*: in order to display accumulated results step by step, we reconstructed subset network that is restructured based on 36 diseases whose final causality exist. Each node in the network indicates disease and is expressed differently depending on disease classification assigned by MeSH. In addition, each node size is proportional to number of connected edges. (a) DAN: Disease association network consists of 1068 disease related edges (grey) and thickness of the edges is proportional to association strength (AS). (b) DPCN: potential causality network consists of 199 potential causality edges (blue) and thickness of such edges is proportional to potential causality strength in Equation (2). (c) DCN: disease causality network consists of 61 causality edges (red) and thickness of such edges is proportional to causality strength in Equation (6)

3.3 Validation on disease causality

Potential causalities in DPCN and causalities in DCN were validated on association strength (AS) in the Section 2.1. Figure 7a shows a result for experiment to compare the number of diseases with PCS in DPCN depending on AS. In total, 1.4 million of disease pairs having AS are sorted in descending order and then such pairs are divided into upper group and lower group depending on size of AS. From the half-top group, we found 85 155 disease pairs having potential causalities, which amounts 64% of total 133 261 potential causalities. This implies that if AS is larger, probability for potential causality is also higher.

On the other hand, Figure 7b shows a result of causality in DCN under the same condition as the previous experiment. It verifies that 58 (95%) disease pairs with causality found in the upper group while in the lower group only 3(5%) disease pairs were found, which amounts to 19 times more than that of the opposite group. Consequently, it can be concluded that causalities are more likely to be found in disease pair with higher AS group.

Table 3 lists the causal and effected disease pairs verified through medical literatures in PubMed. Among the 61 causalities, 16 disease pairs were able to be found in the literature survey. The last column of the table provides, PMID, PMCID, book and other sources. The remaining 45 pairs are undiscovered yet and requires further research. However, these may become prospective candidates for causal and effected disease pairs. The full list of 61 disease pairs are provided in the Supplementary Table S1.

The following are typical phrases from literatures for the causal and effected disease pairs that we found.

3.3.1 Hypertension as a predisposing factor of febrile seizures

Hypertension, which is one of the cardiovascular diseases, is a factor (a pre-disposing factor) to cause febrile seizures which is disease of nervous system (Wallace and Farrell, 2004).

3.3.2 Hypertension as cataract risk factors

Hypertension is verified to be a critical factor to cause cataract in black population (Leske *et al.*, 1999).

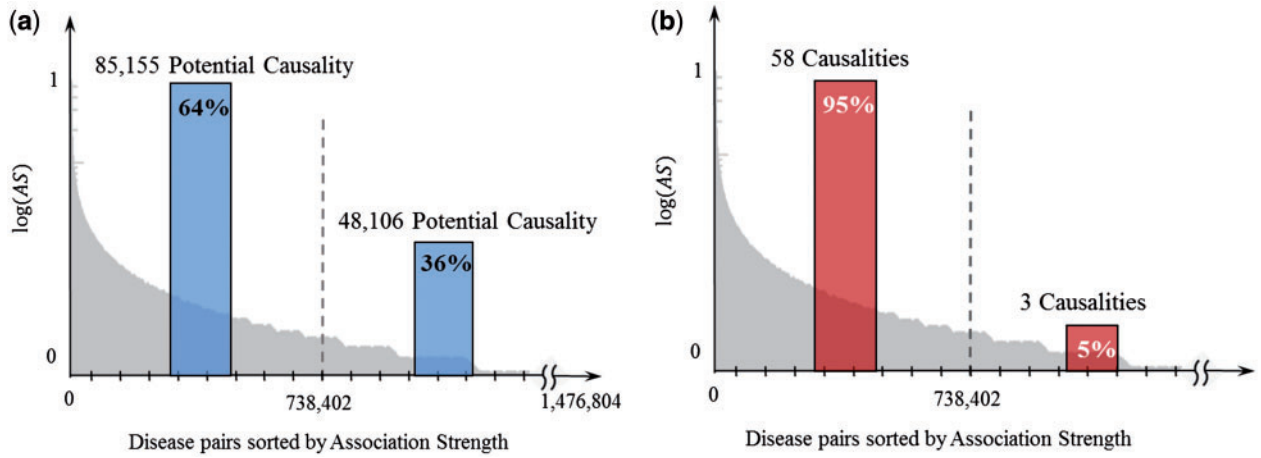

Fig.7. *Validation on disease causality*: (a) shows a result for experiment to compare the number of diseases with PCS in DPCN depending on AS. About 1.4 million of disease pairs having AS are sorted in descending order and then such pairs are divided into upper group and lower group depending on size of AS. From the half-top group, we found 85 155 disease pairs having potential causalities, which amounts 64% of total 133 261 potential causalities and (b) shows a result of causality in DCN under the same condition as the previous experiment. It verifies that 58 (95%) disease pairs with causality found in the upper group while in the lower group only 3 (5%) disease pairs were found

Table 3. Validation for causal and effected disease pairs

	Causal disease	Effected disease	Validation
1	Hypertension	Seizures, Febrile	Epilepsy in Children (2004)
2	Hypertension	Cataract	PMID: 9917778
3	Hypertension	Macular Degeneration	"Macular Degeneration (AMD): Causes, Symptoms and Treatments" Nordqvist
4	Hypertension	Pituitary Neoplasms	PMID: 6551411 Fode *et al.*, 1983
5	Hypertension	Osteopetrosis	PMCID: 1006447
6	Hypertension	Arthritis, Rheumatoid	PMCID: 1006447
7	Atrial Fibrillation	Alzheimer Disease	PMCID: 3289545
8	Lowe syndrome	Huntington Disease	PMCID: 1526415
9	Lowe syndrome	Cerebral Palsy	PMCID: 1526415
10	Tuberous Sclerosis	Huntington Disease	Tuberous sclerosis complex (1999)
11	Tuberous Sclerosis	Parkinson Disease	Gomez *et al.*, 1999
12	Tuberous Sclerosis	Cerebral Palsy	
13	Tuberous Sclerosis	Spinocerebellar Degenerations	
14	Tuberous Sclerosis	Dystonic Disorders	
15	Tuberous Sclerosis	Amyotrophic Lateral Sclerosis	
16	Tuberous Sclerosis	Macular Degeneration	Tuberous sclerosis (2008) Curatolo *et al.*, 2008

3.3.3 Hypertension is some more factors which may contribute to the risk of macular degeneration

People who suffer from hypertension are more at risk of developing macular degeneration (AMD) (Lawrence, 1975).

3.3.4 Atrial fibrillation and risk of dementia, a prospective cohort study

We investigated whether atrial fibrillation is associated with increased risk of incident dementia or Alzheimer disease, beyond its effect on stroke (Dublin *et al.*, 2011).

3.3.5 Lowe syndrome affecting nervous system diseases kind of Huntington disease or cerebral palsy

Lowe syndrome (the oculocerebrorenal syndrome of Lowe, OCRL) is a multisystem disorder characterized by anomalies affecting the eye, the nervous system and the kidney (Loi, 2006).

4 Conclusions

In this study, we proposed a new approach to define causality between diseases. In order to find causality, three different networks were constructed step by step, representing disease association, potential causality and causality, respectively. For each of network, a new method was proposed to extract information from different sources of data crossing disease–gene relations, prevalence and comorbidity frequencies in clinical data, disease–metabolic pathways.

Here are noticeable points in this study:

a. We proposed a systematic framework for network construction method to support ultimate goal to define causality between diseases. The proposed method defined potential causality (DPCN) using prevalence and comorbidity information for disease pairs whose association (DAN) is defined through disease–protein information. In addition, it also confirmed a causality (DCN) for disease pair, whose potential causality exists, by extracting directional information from metabolic pathway. Result of DCN, which is defined by step-by-step analysis for gene information, clinical information and metabolic pathway information, is able to provide more reliable information than when it exist a single entity.
b. The method proposed by this study is a result of real fusion between molecular biology that researches metabolic pathway and medical division that cover patient with certain diseases. Although there were few attempts to define metabolic pathway systematically, it was limited to proposal of method because of excessive complication (Schuster *et al.*, 1999, 2000). In the meantime, although many conventional studies recognized importance of metabolic pathway analysis to define disease association, it is insufficient to reflect directional information interacting genes in cells systematically (Li and Agarwal, 2009). However, this study suggested a systematic method to define causality based on principles by extracting influence of molecular biological flow between diseases from association of genes appeared in metabolic pathway and relations between diseases and metabolic pathway.
c. In addition, this study positively verified that there is a close relation between strength of association and causality. This study also verified that potential causality and causality in top 50% group of higher strength of association is approximately 2 times and 19 times more likely higher, respectively, than those of bottom group.

Further researches following this study could be advanced to two areas: first, as each network consists of different data set, there are many cases to omit information inevitably compared to independent cases when results are accumulated. As this case is not indicating there is not causality but such causality is likely to exist, it is expected that more diversified causalities are clarified by updating data in the future. Second, although causalities that are clarified by the proposed study are verified by medical literatures, not-verified causalities are very likely to be ones which are not verified even by clinical method. Therefore, these results need to be maintained so that they can be used for future Cohort researches. Also, this study remains comparison of our results with a gold standard such as Bradford Hill criteria. It is regarded as a representative causality model in epidemiology (Hill, 1965), which is precise medical discovery but needs demanding time and effort into its cohort study. Note that the proposed study is compatible and complement with such studies. If guided by our causality results, it will reduce a significant amount of time and effort required for conventional approaches.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP; No. 2015R1D1A1A01057178/2012-0000994 to H.J.S.) and the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIP; No.2015R1A5A7037630 to J.H.K.).

Conflict of Interest: none declared.

References

Curatolo,P. *et al.* (2008) Tuberous sclerosis. *Lancet*, **372**, 657–668.

Davis,D.A. and Chawla,N.V. (2011) Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. *PLoS One*, **6**, e22670.

Dublin,S. *et al.* (2011) Atrial fibrillation and risk of dementia: a prospective cohort study. *J. Am. Geriatr. Soc.*, **59**, 1369–1375.

Fode,N.C. *et al.* (1983) Pituitary tumors and hypertension: implications for neurosurgical nurses. *J. Neurosci. Nurs.*, **15**, 33–35.

Goh,K.I. *et al.* (2007) The human disease network. *Proc. Natl. Acad. Sci. USA*, **104**, 8685–8690.

Gomez,M.R. *et al.* (1999). *Tuberous Sclerosis Complex*: Oxford University Press, Oxford.

Hemingway,H. and Marmot,M. (1999) Evidence based cardiology-Psychosocial factors in the aetiology and prognosis of coronary heart disease: systematic review of prospective cohort studies. *BMJ*, **318**, 1460–1467.

Hidalgo,C.A. *et al.* (2009) A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.*, **5**, e1000353.

Hill,A.B. (1965). The environment and disease: association or causation? *Proc. R. Soc. Med.*, **58**, 295.

Kukull,W.A. *et al.* (2002) Dementia and Alzheimer disease incidence: a prospective cohort study. *Arch. Neurol.*, **59**, 1737–1746.

Lawrence,J. (1975) Hypertension in relation to musculoskeletal disorders. *Ann. Rheum. Dis.*, **34**, 451–456.

Leske,M.C. *et al.* (1999) Diabetes, hypertension, and central obesity as cataract risk factors in a black population: The Barbados Eye Study. *Ophthalmology*, **106**, 35–41.

Li,Y. and Agarwal,P. (2009) A pathway-based view of human diseases and disease relationships. *PLoS One*, **4**, e4346.

Lillioja,S. *et al.* (1993) Insulin resistance and insulin secretory dysfunction as precursors of non-insulin-dependent diabetes mellitus: prospective studies of Pima Indians. *N. Engl. J. Med.*, **329**, 1988–1992.

Loi,M. (2006) Lowe syndrome. *Orphanet J. Rare Dis*., **1**, 16.

McDonald,G.B. *et al*. (1993) Veno-occlusive disease of the liver and multiorgan failure after bone marrow transplantation: a cohort study of 355 patients. *Ann. Intern. Med*., **118**, 255–267.

McNutt,L.A. *et al*. (2003) Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am. J. Epidemiol*., **157**, 940–943.

Nordqvist,C. (2015, December 22). *Macular degeneration (AMD): causes, symptoms and treatments*. Medical News Today. http://www.medicalnewstoday.com/articles/152105.php.

Saltiel,A.R. and Olefsky,J.M. (1996) Thiazolidinediones in the treatment of insulin resistance and type II diabetes. *Diabetes*, **45**, 1661–1669.

Schuster,S. *et al*. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol*., **17**, 53–60.

Schuster,S. *et al*. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol*., **18**, 326–332.

Wallace,S.J. and Farrell,K. (2004). *Epilepsy in Children, 2nd edn*. CRC Press, Boca Raton.

Xu,R. *et al*. (2014) dRiskKB: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text. *BMC Bioinformatics*, **15**, 105.

Zhang,J. and Kai,F.Y. (1998) What's the relative risk?: a method of correcting the odds ratio in cohort studies of common outcomes. *JAMA*, **280**, 1690–1691.

Zhang,X. *et al*. (2011) The expanded human disease network combining protein–protein interaction information. *Eur. J. Hum. Genet*., **19**, 783–788.