BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 16 2013, pages 2017–2023
doi:10.1093/bioinformatics/btt327

Data and text mining

Advance Access publication June 6, 2013

Inferring disease association using clinical factors in a combinatorial manner and their use in drug repositioning

Jinmyung Jung* and Doheon Lee*

Department of Bio and Brain Engineering, KAIST, 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Republic of Korea

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Complex physiological relationships exist among human diseases. Thus, the identification of disease associations could provide new methods of disease care and diagnosis. To this end, numerous studies have investigated disease associations. However, combinatorial effect of physiological factors, which is the main characteristic of biological systems, has not been considered in most previous studies.

Results: In this study, we inferred disease associations with a novel approach that considered disease-related clinical factors in combinatorial ways by using the National Health and Nutrition Examination Survey data, and the results have been shown as disease networks. Here, the FP-growth algorithm, an association rule mining algorithm, was used to generate a clinical attribute combination profile of each disease. In addition, we characterized the 22 clinical risk attribute combinations frequently discovered from the 26 diseases in this study. Furthermore, we validated that the results of this study have great potential for drug repositioning and outperform other existing disease networks in this regard. Finally, we suggest a few disease pairs as new candidates for drug repositioning and provide the evidence of their associations from the literature.

Contact: dhlee@kaist.ac.kr or jmjung.kr@gmail.com

Supplementary information: Supplementary data are available at the *Bioinformatics* online.

Received on January 14, 2013; revised on May 28, 2013; accepted on June 2, 2013

1 INTRODUCTION

Complex physiological relationships exist among human diseases. Thus, the identification of disease similarity has been used to study etiology and pathogenesis of similar diseases (Kalaria, 2002), as well as have provided novel ways for disease diagnosis, drug repositioning (Barabási *et al*., 2011; Goh *et al*., 2007; Suthram *et al*., 2010). Therefore, many studies to elucidate disease associations have emerged, and the results of which have been presented as disease networks (DNs). Previous studies of DNs can be classified into two categories based on the used data, molecular data and electronic health record (EHR) data.

From a molecular perspective, Goh *et al*. constructed DNs that were linked if they shared one or more disease genes (Goh *et al*., 2007). Following Goh's study, functional module–based studies by using molecular data have also appeared to clarify disease associations. Lee *et al*. created DNs that were linked if the enzymes related to diseases shared metabolic reactions (Lee *et al*., 2008). In addition, Suthram *et al*. (2010) also built DNs by an integrated analysis of disease-related mRNA expression data and the human protein interaction networks. Here, two diseases were linked if they had similar profiles of 4620 functional modules that were described by differently expressed genes of the disease. Park *et al*. (2011), more recently, published the article of disease associations by considering both disease-related proteins and their localization in a cell.

The other approach for constructing DNs based on EHR data have also emerged. EHR data have usually been used for obtaining comorbidity, which implies the presence of one or more diseases in addition to a primary disease that the patient has. For example, Hidalgo *et al*. (2009) introduced DNs obtained from the disease history of more than 30 million patients. Roque *et al*. (2011) also presented networks from phenotype information including both structured and unstructured electronic patient records, such as free-text diagnosis reports. Furthermore, Holmes *et al*. (2011) introduced ADAMS for discovering disease associations using multiple sources, such as PubMed articles, discharge summaries and so on.

However, previous studies overlooked the combinatorial effect, which is one of main biological characteristics. Combinatorial effects exist in biological system, and biological factors influence a disease in combinatorial rather than individual ways.

As evidence, studies for combinatorial effects in biology and physiology have recently appeared. In molecular biology field, Knijnenburg *et al*. (2009) presented that most *Saccharomyces cerevisiae* genes were shown to be influenced by combinatorial effects of cultivation parameters. These combinatorial effects led to higher explained variance of the gene expression patterns. Liu *et al*. (2011) showed that the combination of rapamycin and lapatinib significantly decreased growth of triple-negative breast cancers. We also explored some clinical studies devoted to identifying the combinatorial effects of diseases (Bruzzi *et al*., 1985; Emmons *et al*., 1994; Gorell *et al*., 2004; Stamler *et al*., 1993). Furthermore, the combinatorial approach has been applied in adverse drug event field. Harpaz *et al*. (2010) extracted multi-item adverse drug events from the Food and Drug Administration's adverse effect reporting system by using association rule mining. All of these studies have shown that combinatorial effects exist in physiological system and are worthy to be regarded to elucidate diseases and disease associations.

*To whom correspondence should be addressed.

The fundamental assumption of previous studies, on the other hand, is that higher association exists between diseases as more disease-related factors are shared individually, not combinatorially. Especially in Goh's study (Goh *et al.*, 2007), individual sharing genes is a measure of similarity. However, even though two diseases share disease genes, the etiological factors of each disease could be totally different due to the combinatorial effects on biological factors.

For example, there are two diseases sharing disease genes, gene A and B. One disease occurs when both gene A and gene B are not working properly; however, the other disease presents when gene A or gene B is not functioning. Then, it is hard to conclude that these two diseases have a perfect association (more explanation in Fig. 1). Thus, when we try to get more precise and accurate association between diseases from a biology perspective, combinatorial effect profiles of them should be analyzed. This explanation can be applied to other molecule-based studies and EHR-based studies in the same manner.

From this background, we present a novel approach that considers disease-related physiological factors in combinatorial ways for inferring disease associations. Here, we decided to use clinical data, rather than molecular data or comorbidity data. Molecular information is incomplete and should be further explored, even though much knowledge has been discovered by development of recent technology and analysis. On the other hand, clinical data provide practical explanations of disease associations and is becoming available to researchers in massive amounts (Holmes *et al.*, 2011).

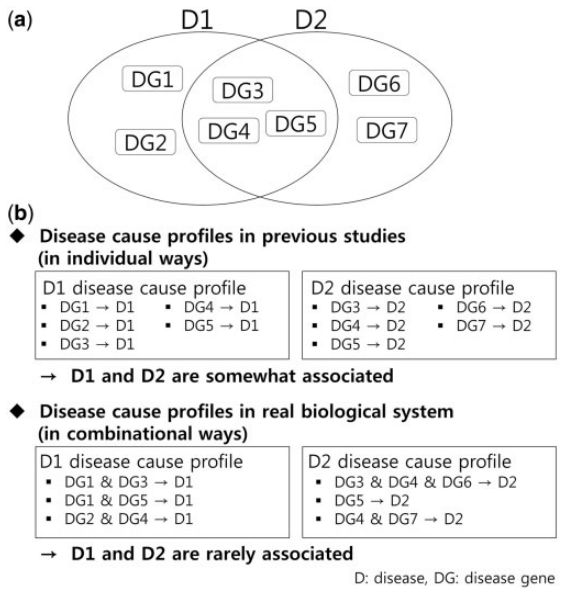

Fig. 1. Possible problems when combinatorial effects are overlooked for inferring disease association. (**a**) Diagram indicating disease genes of two diseases (D1, D2). (**b**) Assumed disease cause profiles in previous studies and real biological system. In Goh's study (Goh *et al.*, 2007), D1 and D2 are considered to be somewhat associated based on the assumption that disease genes affect each disease individually (upper half). However, each disease cause profile could be different in real biological system because of combinatorial effects, and two diseases might be rarely associated (lower half)

The National Health and Nutrition Examination Survey (NHANES; http://www.cdc.gov/nchs/nhanes.htm), a major clinical survey of the National Center for Health Statistics, was used as primary data for this study. Molecule and comorbidity data were incorporated to evaluate the results of this study in evaluation part as well.

In this article, we constructed the DNs from the information of disease-related clinical attribute combinations (Fig. 2). In addition, we generated the 22 clinical attribute combinations frequently discovered from the 26 diseases concerned in this study, named as 'Clinical Risk Attribute Combinations (CRACOMs)'. In the evaluation section, we showed that the results of this study have great potential to be used for drug repositioning and outperform other existing DNs in this regard. Furthermore, we investigated a few disease pairs as candidate disease for drug repositioning and present some valuable evidence of the associations.

2 METHODS

2.1 Data introduction

NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the USA. NHANES is a major survey of National Center for Health Statistics, which is a part of the Centers for Disease Control and Prevention. Owing to some characteristics of this survey that are various data, including both interviews and physical examinations data, and own sequence number to every sample person, many researchers have used it for epidemiological studies and health sciences research (Healy *et al.*, 2011; Looker *et al.*, 2010).

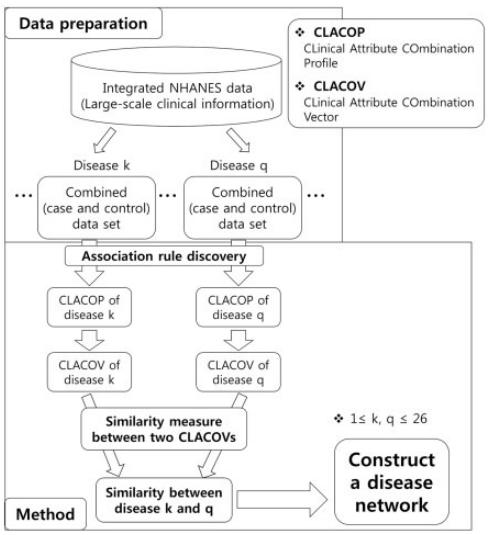

Fig. 2. Overall procedures of this study. In the data preparation part, we organized combined datasets, each of which is composed of both the case and control data of each disease, from integrated NHANES data. In the method part, we obtained disease-specific CLACOPs, each of which turned out to affect each disease, by applying an association rule mining algorithm to the combined datasets. From these CLACOPs, we generated CLACOVs and calculated similarities between CLACOV pairs for the purpose of deducing similarities between disease pairs. We considered 26 diseases in this study

2.2 Data preparation

For the purpose of this research, we obtained large-scale clinical information by integrating from 1999 to 2010 Questionnaire and Laboratory data from NHANES, resulting in the 62 160 instances and 3812 attributes (Fig. 3a). From the integrated NHANES dataset, we generated the case dataset of each disease, which was composed of instances that had the disease in the past. Among many diseases included in NHANES data, we selected 26 diseases, each of which has more than 450 instances in a case dataset. Then, we applied a filtering algorithm to each case dataset so as to filter out instances and attributes having more missing values than the predefined threshold. After filtering, each case dataset has its own instances and attributes (Fig. 3b). To analyze all of 26 diseases impartially, we decided to use the only 73 attributes shared among 26 disease case dataset (Supplementary Table S1). Next, we appended a control dataset, which contains instances that are not related to the disease, to the corresponding case dataset to construct a combined dataset (Fig. 3c).

After constructing combined dataset, we transformed every value in the dataset into a binary value based on reference range of each attribute (Supplementary Table S1) (Iverson *et al.*, 2007). In this transformation, we mapped every value to either 1 (if a value does NOT lie in reference range) or 0 (if a value lies in reference range). As the last step, Random imputation overall method was applied to all combined dataset to impute missing values (Kalton and Kasprzyk, 1982). These datasets were used as an input of an association rule mining algorithm

2.3 Clinical attribute combination profile

As we mentioned in the introduction part, a disease occurs by plenty of causes, and these causes influence a disease in combinatorial ways rather than in independent ways. Therefore, analysis of clinical factors in combinatorial ways is necessary to research disease associations. To this end, for each of the 26 diseases, we obtained Clinical Attribute Combination Profile (CLACOP) composed of a set of disease-specific clinical attributes that are considered to influence the disease in combinatorial ways.

To generate a CLACOP of each disease, the FP-growth algorithm, one of the efficient association rule mining algorithms, was applied to each combined dataset. The FP-growth algorithm needs much lower computing power than the Apriori algorithm, as it is applied to binary dataset, which is a characteristic of the preprocessed combined dataset.

When we use FP-growth algorithm, three thresholds such as support, confidence and max number of attributes should be determined. However, it was ambiguous and not easy to set above three thresholds because the final results, both disease similarities (correlations) and associations, could be different according to them. For setting the appropriate thresholds, we experimented on the whole procedures (refer to section 2.3, 2.4, 3.1) with 17 sets of different thresholds and generated disease correlation ranks for each threshold (Supplementary Table S2). Then, we measured their Spearman's rank correlation coefficients in pair-wise manner, resulting in subtle difference among them (Supplementary Table S3). In the end, we selected the thresholds (support: 0.05, confidence: 0.7, max number of attributes: 5) to be used in the main experiment, having the highest Spearman's rank correlation coefficient compared with the average disease correlation rank (Supplementary Table S4).

Therefore, WEKA program (Hall *et al.*, 2009) was used for running the FP-growth algorithm with three thresholds (support: 0.05, confidence: 0.7, max number of attributes: 5) to each combined dataset (Fig. 4a). Among the rules that the FP-growth algorithm have generated, we selected rules having the corresponding disease attribute only in consequent part. Here, the antecedent of the selected rules was postulated as a clinical attribute combination that affects the corresponding disease in combinatorial ways (Fig. 4b). Then consequently, we generated a set of clinical attribute combinations (CLACOP) for the disease. By performing the same procedure to 26 diseases, finally, we obtained 26 CLACOPs (Fig. 4a).

2.4 Disease similarity calculation

For identifying disease similarity in quantitative, we transformed 26 CLACOPs to a Clinical Attribute Combination Vectors (CLACOVs) of the disease. Then, we applied the cosine similarity method to a pair

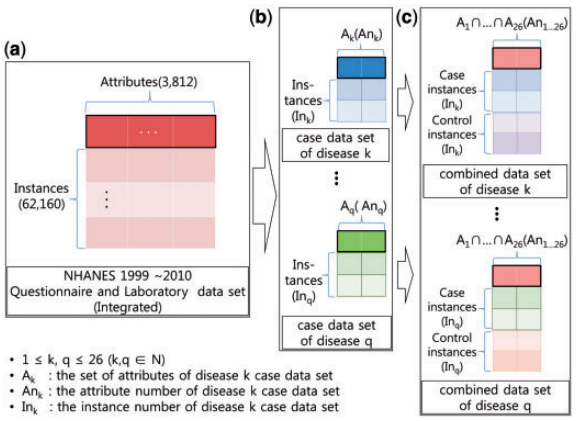

Fig. 3. Data preparation by preprocessing the integrated NHANES data (the color figure in Supplementary Fig. S10). (**a**) Integrated from 1999 to 2010 NHANES data. Large-scale clinical information was obtained by integrating from 1999 to 2010 Questionnaire and Laboratory data from NHANES. (**b**) Case dataset of each disease. After applying a filtering algorithm, each case dataset has its own instances (the instance number of disease k: Ink) and attributes (Ak). (**c**) Combined dataset of each disease. The shared attributes among 26 filtered case dataset (A1∩ . . . ∩A26) were selected and used. The number of shared attributes is 73 (An1 . . . 26) (Supplementary Table S1). Finally, we appended control dataset to case dataset to construct combined dataset

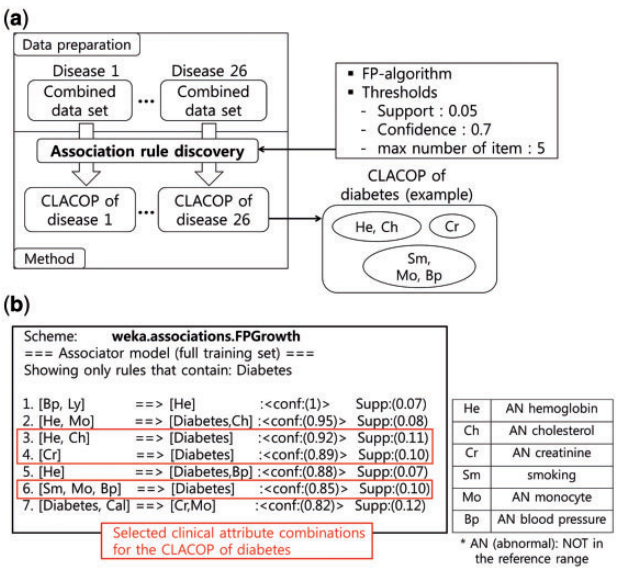

Fig. 4. (**a**) Procedures for generating CLACOP. (**b**) The FP-growth algorithm results of diabetes dataset (imaginary example). Among the rules that the FP-growth algorithm have generated, we only included the rules that had the only corresponding disease attribute in consequent part to the CLACOP of a disease

of CLACOVs for calculating similarity of two diseases. Here, we decided to use the cosine similarity method because the method has been numerously used in text mining area when determining how two records are similar (Bilenko and Mooney, 2003; Larsen and Aone, 1999), which is analogous to this procedure. Cosine similarity is given as

$$\text{Similarity} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

When we generated a CLACOV, we assigned the confidence of the rule to the corresponding clinical attribute combination (Fig. 5a). As a result, we generated 26 CLACOVs by using 26 CLACOPs and their confidence scores. Then, the cosine similarity method was applied to a pair of CLACOVs, and they were regarded as similarities between two diseases (Fig. 5b).

3 RESULTS

3.1 The disease correlation matrix

For calculating more reliable and unbiased disease similarities, overall procedures were repeated five times with different control dataset, which were selected randomly from each control data pool (Supplementary Figure S1). In the end, the CLACOV-sharing disease correlation (similarity) matrix was constructed with the average similarities. We visualized it in Figure 6a, and the correlations were described in Supplementary Table S5.

To analyze and compare the disease correlation matrix (DCM) in detail, we also constructed three other DCMs based on different entity such as gene, single-nucleotide polymorphism (SNP) and patient data. We applied the individual approach (not the combinatorial approach), which is the conventional approach for DNs, for generating those three DCM. To this end, we calculated the similarity between a pair of diseases, indicating how many entities (gene, SNP or patient) are shared. For example, for generating gene-sharing DCM, we generated a list of genes known to be associated with each disease, and the disease similarity (correlation) was calculated based on how many genes are shared between a pair of diseases. The similarity is defined as

$$\text{Similarity} = \frac{\text{N}(g_k \cap g_q)}{\sqrt{\text{N}(g_k)} \times \sqrt{\text{N}(g_q)}} \quad (2)$$

where $\text{N}(g_k)$ is the number of genes used to disease k, and $\text{N}(g_k \cap g_q)$ is the number of genes used to both disease k and disease q. SNP-sharing, patient-sharing DCM were also generated with the same way used for gene-sharing DCM (Supplementary Fig. S2). Their correlations were specified in Supplementary Table S6, and

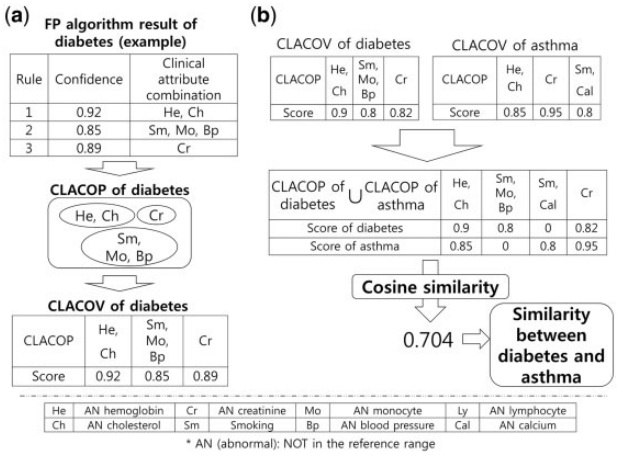

Fig. 5. (**a**) Procedures for generating CLACOV. CLACOV was generated by using both CLACOP and their scores that are obtained from confidences of the rules. (**b**) Disease similarity calculation between two CLACOVs. The union of two CLACOPs was created in every disease pair similarity calculation so that the cosine similarity method was able to be applied to a pair of CLACOVs, and those results were regarded as similarities between two diseases. Here, we presented imaginary examples for making it easy to understand

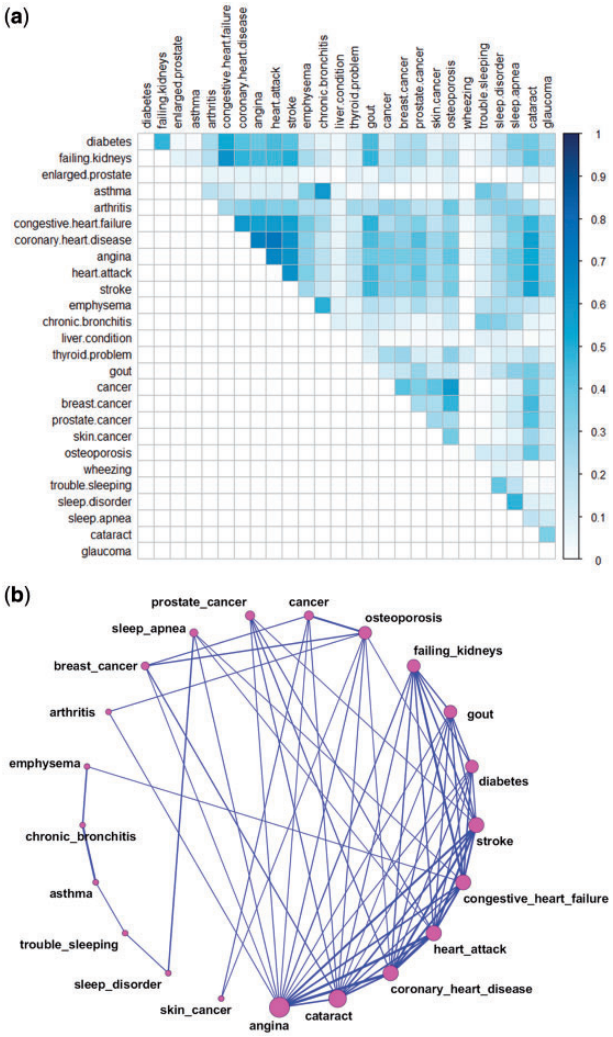

Fig. 6. (**a**) CLACOV-sharing DCM. Darkness of the blue color corresponds to correlation. (**b**) The DNs of this study. We only figured correlations that are more than 0.35. The width of edge is proportional to the square of correlation, and the node size corresponds to the degree of the node. The DNs consisting of significant associations compared with random sets were constructed in Supplementary Figure S4. (The color figure has been provided in Supplementary Fig. S11.)

the used databases for each entity are specified in Supplementary Figure S3.

When comparing the DCM of this study with the others, especially with gene-sharing DCM, we see many expected disease correlations such as the similarities among heart attack, angina and congestive heart failure. In addition, well-known correlation between failing kidneys and heart diseases was also discovered. We also see some unexpected correlations such as the gout and heart disease correlation as well as the cataract and heart failure correlation.

3.2 The disease networks

For assigning significance to the obtained disease similarities (correlations), we generated a background distribution of disease similarities by random chance for each pair of diseases. To this end, for each disease, we prepared a randomized combined dataset, in which we shuffled case and control instances randomly. Then, we tried to generate CLACOP with the same threshold used in this study (support: 0.05, confidence: 0.7, max number of attributes: 5). However, few clinical attribute combinations for each disease resulted that few disease pairs have non-zero similarities. That is because thresholds, especially confidence, were too strict to generate any clinical attribute combinations from randomly shuffled dataset. Therefore, we applied lower confidence (0.55) for obtaining background distribution. Because the confidence is directly used for calculating a disease similarity, adjusting confidence than other thresholds could give more reasonable results. Finally, we created background distributions for each pair of diseases by repeating the whole process 100 times (Supplementary Table S7).

Next, we selected only those disease pairs having less *P*-value–like significance score than 0.0001 (FDR-like score = 0.01%), resulting in 260 significant disease associations (Supplementary Table S7), and we constructed the final DNs by using those associations. (Here, the terms such as '*P*-value–like' or 'FDR-like' have been used because different confidence thresholds were applied to compute the foreground and background disease similarities.) We, however, decided to visualize the networks in supplementary material (Supplementary Fig. S4), as it shows a complicated picture. Instead, the DNs consisting of edges that have higher correlations than 0.35 were built and shown in Figure 6b.

3.3 Clinical risk attribute combination

Furthermore, we analyzed the results to discover clinical attribute combinations associated with various types of diseases. We postulated that there exist the clinical attribute combinations that are prevalent in various diseases because diseases are associated with each other in some way. By investigating those combinations, causes of comorbidity or pathogenic patterns of diseases can be analyzed. Thus, we characterized CRACOMs that are clinical attribute combinations appeared in 17 or more diseases among the 26 diseases, resulting in 22 CRACOMs (Supplementary Table S8). The CRACOMs would be worth elucidating further for disease etiology and comorbidity.

4 EVALUATION

4.1 Potential for drug repositioning

Knowledge of a disease similarity based on combinatorial clinical factors could be applied to find new uses for existing drugs. Similar diseases share similar symptoms or clinical factors and could be potentially cured by similar drugs. To this end, we checked the potential that this study is able to be used in drug repositioning by comparing with drug-sharing DNs, which were built based on drug-sharing DCM.

For constructing drug-sharing DCM, we used the same method that was used to construct gene-sharing DCM (2). In detail, we generated a list of drugs known to be associated with each disease, and the disease similarity was calculated based on how many drugs were shared between a pair of diseases. Finally, we could obtain drug-sharing DCM, resulting in drug-sharing DNs.

Drug-sharing DNs have been regarded as reference DNs in this evaluation. We constructed two reference DNs according to resources (Supplementary Fig. S5). The first reference was built based on DrugBank, PharmGKB and TTD database, all of which contains information of approved drugs. The second reference was built based on clinicaltrials.gov database, which contains information of drugs that are not approved yet but in the process of trials on patients (Supplementary Fig. S6 and Table S9).

After constructing the reference DNs, we extracted 30 disease pairs in CLACOV-sharing DNs that have higher similarities than others (top 30 ranked disease pairs), and we compared those pairs to top 30 ranked disease pairs of two reference DNs. It resulted in that 18 disease pairs and 15 diseases pairs shared, respectively, and for both, $P < 0.0001$ (Supplementary Fig. S7). Thus, we concluded that the result of this study is worthy to be used for drug repositioning. Additionally, we closely examined 18 disease pairs that shared with the first reference networks (Supplementary Fig. S8). In results, some disease pairs are biased toward several disease classes such as cardiovascular disease and respiratory disease. The reason for the bias, which we have concluded, is that NHANES data mainly contain the information of the prevalent diseases such as diabetes, stroke and heart attack. In addition, reducing diseases by the preprocessing, which filtered out the diseases having more missing values than the threshold, could be one of the causes. In the future work, we are planning to construct DNs with increased disease coverage by using more various data.

4.2 Outperformance in drug repositioning field

There exist many disease similarity studies based on other data, such as gene, SNP and comorbidity, and they might provide better information to drug repositioning research than this study. Thus, we compared the performance of CLACOV-sharing DNs with other DNs to see how better performance we could offer.

To this end, we built three DNs based on gene-sharing DCM, SNP-sharing DCM and patient-sharing DCM (comorbidity), which were generated in section 3.1. For each DN (CLACOP-sharing, gene-sharing, SNP-sharing and patient-sharing DNs), we measured how many disease pairs share with reference DNs

among top-k ranked disease pairs while k was increasing from 1 to 50 (Supplementary Fig. S9). This evaluation shows that this study outperformed other approaches (Fig. 7). Therefore, we believe that CLACOP-sharing DNs potentially give useful information to drug repositioning field.

The combinatorial approach in this study could give one explanation of this good performance. When a drug is taken to the body, the drug gives rise to a few effects including an expected effect because the biology system of the body is intricate. Then, the combinatorial action of them (biological system basically functions in a combinatorial manner as mentioned in introduction part) leads to a phenotypic effect such as cure of disease (Fig. 8a). Thus, if two diseases share the combinatorial profile that is highly correlated to the corresponding drug effects, the two diseases have more possibility to be drug-repositioned than other diseases (Fig. 8b).

In addition, utilization of clinical data itself could be an advantage in drug research field. Other conventional approaches for drug repositioning, such as target-based approach (Keiser *et al.*, 2009) and effect-based approach (Sirota *et al.*, 2011), have the same limitation that is the translational problem, which means that a drug does not work in clinical trials even though it works well in an *in vitro* or a mouse experiment. This study, however, can complement the translational problem because clinical data were directly used for calculating disease associations and inferring drug repositioning candidate diseases.

5 DISCUSSION

We further explored disease pairs that are able to be new candidates for drug repositioning. As a result, we found that the gout and heart diseases pairs were positioned at high ranks in this study, but drugs hardly shared for the disease pairs yet (Table 1). To verify the potential for new candidates of drug repositioning, we investigated relationships between gout and heart diseases from literature, and we could obtain some evidence that they are closely associated. Choi and Curhan (2007) and De Vera *et al.* (2010) claimed that men and women with gout have an increased risk for heart diseases, such as myocardial infarction and coronary heart disease. Moreover, Krishnan *et al.* (2006) also revealed that gout is a risk factor of myocardial infarction, using multivariable regression models. In Krishnan's article, they referred that further research will be needed for the exact links between two diseases, and the combinatorial approach of this study could give some clues for it.

We also disclosed disease pairs of cataract and heart diseases that has great potential for new drug repositioning candidates (Table 1). To understand the results further, we surveyed articles that support the observed disease association. Theodoropoulou *et al.* (2011) reported that coronary heart disease is one of the statistically significant risk factor of cataract. In addition, Younan et al. (2003) provided some evidence supporting a relationship between cardiovascular diseases and incident cataract.

6 CONCLUSION

The combinatorial approach is highly necessary to explain disease associations because the real diseases occur in combinatorial ways. In this study, we inferred disease associations with a novel approach that considered disease-related physiological factors in combinatorial ways by using NHANES data. This study will give good explanations for understanding disease associations, and it will also provide inspiration to combinatorial approach studies not only in the disease related field but also in other fields having combinatorial characteristic. Furthermore, we generated disease specific CLACOPs as well as CRACOMs, which will give a few clues for disease prevention and etiology. Those profiles will be investigated in the future.

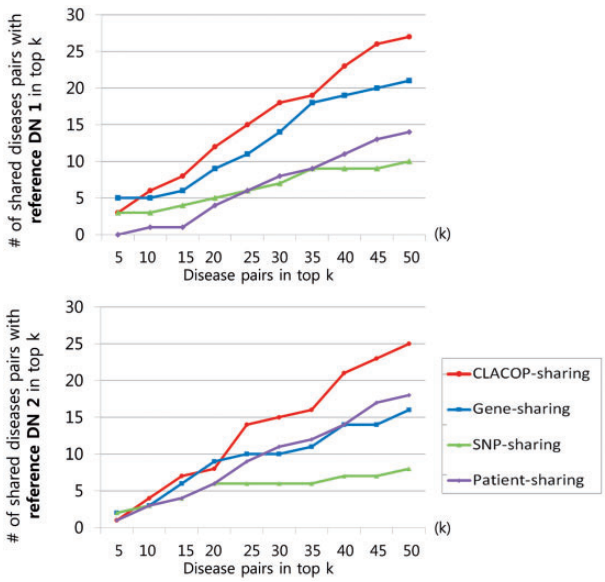

Fig. 7. The evaluation results (the color figure in Supplementary Fig. S12). This results show that the networks of this study (CLACOP-sharing DN) performed better than other networks at almost every top k, when compared with both reference networks 1 (approved drug-sharing DN) and reference networks 2 (trial drug-sharing DN)

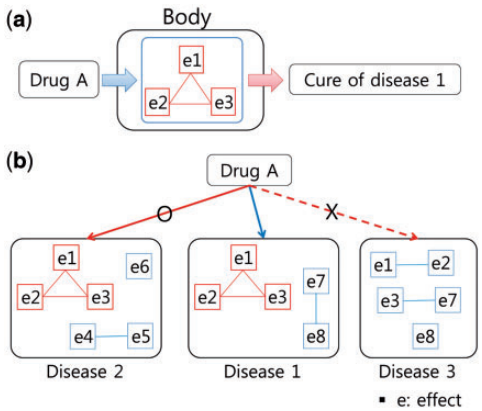

Fig. 8. (**a**) The combinatorial effects of drug A to disease 1. This figure describes the combinatorial effects of drug A resulting in cure of disease 1. (**b**) Drug repositioning of the drug A for disease 2. Even though diseases 1 and 3 share more individual effects than diseases 1 and 2, the drug A is repositioned to disease 2, as the combinatorial effects caused by the drug A exists in both diseases 1 and 2, not disease 3. (The color figure has been specified in Supplementary Fig. S13)

Table 1. New disease candidates for drug repositioning

Disease pairs	Similarity		
	CLACOP-sharing DN	Ref. 1 DN	Ref. 2 DN
Cataract and CHD	0.564	0.0	0.0
Gout and CHF	0.481	0.0	0.003
Cataract and CHF	0.476	0.0	0.0
Gout and CHD	0.450	0.0	0.0

Note: Four disease pairs in this table have been revealed as candidates for drug repositioning. Those pairs had relatively high similarities in this study but relatively low similarities in the reference DN, indicating that they have not been used for drug 55 repositioning yet in spite of the potential. CHF: congestive heart failure, CHD: coronary heart disease.

We also validated the potential of this study to be used in the drug repositioning field, and showed that this study outperformed other approaches in this regard. As a result, a few disease pairs such as the gout and congestive heart failure, and the cataract and coronary heart disease, were suggested as the candidate diseases for drug repositioning.

We will investigate these drug repositioning candidate diseases on hereafter studies. In addition, when molecular data and their combinatorial analysis are incorporated, further improvements could be achieved in elucidating disease associations because molecular data provide the information that physiological data cannot cover. This further work also will be investigated in the future.

Funding: Biomedical Technology Development program (NRF-2012M3A9C4048758); Basic Research Laboratory (2009-0086964) of the Ministry of Science, ICT and Future Planning through the National Research Foundation of Korea.

Conflict of Interest: none declared.

REFERENCES

Barabási,A.L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.

Bilenko,M. and Mooney,R.J. (2003) Adaptive duplicate detection using learnable string similarity measures. In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, Washington, DC. pp. 39–48.

Bruzzi,P. *et al.* (1985) Estimating the population attributable risk for multiple risk factors using case-control data. *Am. J. Epidemiol.*, **122**, 904–914.

Choi,H.K. and Curhan,G. (2007) Independent impact of gout on mortality and risk for coronary heart disease. *Circulation*, **116**, 894–900.

De Vera,M.A. *et al.* (2010) Independent impact of gout on the risk of acute myocardial infarction among elderly women: a population-based study. *Ann. Rheum. Dis.*, **69**, 1162–1164.

Emmons,K.M. *et al.* (1994) Mechanisms in multiple risk factor interventions: smoking, physical-activity, and dietary-fat intake among manufacturing workers. *Prev. Med.*, **23**, 481–489.

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

Gorell,J.M. *et al.* (2004) Multiple risk factors for Parkinson's disease. *J. Neurol. Sci.*, **217**, 169–174.

Hall,M. *et al.* (2009) The WEKA data mining software: an update. *ACM SIGKDD Explor. Newslett.*, **11**, 10–18.

Harpaz,R. *et al.* (2010) Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics*, **11**, S7.

Healy,G.N. *et al.* (2011) Sedentary time and cardio-metabolic biomarkers in US adults: NHANES 2003–06. *Eur. Heart J.*, **32**, 590–597.

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

Holmes,A.B. *et al.* (2011) Discovering disease associations by integrating electronic clinical data and medical literature. *PloS One*, **6**, e21132.

Iverson,C. *et al.* (2007) *AMA Manual of Style: A Guide for Authors and Editors*. Oxford, New York.

Kalaria,R. (2002) Similarities between Alzheimer's disease and vascular dementia. *J. Neurol. Sci.*, **203**, 29–34.

Kalton,G. and Kasprzyk,D. (1982) *Imputing for Missing Survey Responses*. Proceedings of the Survey Research Methods Section, 22-31.

Keiser,M.J. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature*, **462**, 175–181.

Knijnenburg,T.A. *et al.* (2009) Combinatorial effects of environmental parameters on transcriptional regulation in *Saccharomyces cerevisiae*: a quantitative analysis of a compendium of chemostat-based transcriptome data. *BMC Genomics*, **10**, 53.

Krishnan,E. *et al.* (2006) Gout and the risk of acute myocardial infarction. *Arthritis Rheum.*, **54**, 2688–2696.

Larsen,B. and Aone,C. (1999) Fast and effective text mining using linear-time document clustering. In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 16–22.

Lee,D.S. *et al.* (2008) The implications of human metabolic network topology for disease comorbidity. *Proc. Natl Acad. Sci. USA*, **105**, 9880.

Liu,T. *et al.* (2011) Combinatorial effects of Lapatinib and rapamycin in triple-negative breast cancer cells. *Mol. Cancer Ther.*, **10**, 1460–1469.

Looker,A.C. *et al.* (2010) Prevalence and trends in low femur bone density among older US adults: NHANES 2005–2006 compared with NHANES III. *J. Bone Miner. Res.*, **25**, 64–71.

Park,S. *et al.* (2011) Protein localization as a principal feature of the etiology and comorbidity of genetic diseases. *Mol. Syst. Biol.*, **7**, 494.

Roque,F.S. *et al.* (2011) Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput. Biol.*, **7**, e1002141.

Sirota,M. *et al.* (2011) Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.*, **3**, 96ra77.

Stamler,J. *et al.* (1993) Diabetes, other risk factors, and 12-yr cardiovascular mortality for men screened in the multiple risk factor intervention trial. *Diabetes Care*, **16**, 434–444.

Suthram,S. *et al.* (2010) Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput. Biol.*, **6**, e1000662.

Theodoropoulou,S. *et al.* (2011) The epidemiology of cataract: a study in Greece. *Acta Ophthalmol.*, **89**, e167–e173.

Younan,C. *et al.* (2003) Cardiovascular disease, vascular risk factors and the incidence of cataract and cataract surgery: the Blue Mountains Eye study. *Ophthalmic Epidemiol.*, **10**, 227–240.