Journal of
Rare Diseases Research
& Treatment

Mini Review

Open Access

Mechanism-based disease similarity

Mehdi B Hamaneh, and Yi-Kuo Yu*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

Article Info

Article Notes
Received: September 07, 2016
Accepted: October 18, 2016

***Correspondence:**
Dr. Yi-Kuo Yu, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
E-mail: yyu@ncbi.nlm.nih.gov

Keywords:
Disease-disease similarity
Human disease network
Disease-gene associations
Protein-protein interaction network
gene-gene association network
gene expression data

ABSTRACT

In recent years several methods have been proposed to assign pairwise mechanism-based similarity scores to human diseases. Despite their differences in approach and performance, these methods work in a somewhat similar manner: first a set of biomolecules (genes, proteins, chemicals, etc.) is associated with each disease, and then a measure is defined to calculate the similarity between the sets assigned to a pair of diseases. Since the similarity score between two diseases is defined based on the underlying molecular processes, a high score may hint at a shared cause, and therefore a similar treatment, for both diseases. This is of great practical importance especially when a rare or newly-discovered disease, for which limited information is available, is found to be related to a disease with a known treatment. Thus, in this mini-review we briefly discuss the recently developed methods for computing mechanism-based disease-disease similarities.

Disease similarity or disease-disease associations can be viewed from different perspectives. Here, by similarity we mean sharing the same underlying molecular processes. Although a strong correlation between having the same symptoms and sharing disease genes has been reported[1], there are many diseases with the same underpinning biological processes that have different symptoms (for some examples see Hamaneh and Yu[2]). For this reason, and due to the limited space, we focus on the functional (mechanism-based) similarity measures and do not cover studies that have used semantic/phenotypic similarities. Mechanism-based methods generally assign a set of genes (and sometimes other biomolecules) to each disease and compare the sets associated with two diseases to find their similarity. As explained in the following paragraphs, depending on the way the sets are associated with diseases, the methods can be classified in the following categories: gene-based, pathway-based, expression-based, network-based, and methods that use (in addition to genes) other biomolecules.

Gene-based methods

Many proposed functional disease relatedness measures have been influenced by the work of Goh and co-workers[3] who created a human disease network (HDN) by linking diseases that share (according to Online Mendelian Inheritance in Man (OMIM[4])) at least one implicated gene. Despite its simplicity, the HDN has been shown to have some predicting power. For example, the authors showed that the diseases in the HDN cluster according to their known classes. Bauer-Mehren *et al.*[5] used the same approach, but employing multiple gene-disease association sources, to create disease-gene, disease-disease and gene-gene association networks and revealed functional modules in complex, environmental, and Mendelian diseases. In another study, the number of genes shared between two diseases was reported to have a statistically significant, although weak, correlation with their comorbidity[6].

Pathway-based methods

Despite its aforementioned successes, the method of Goh *et al.* is not able to uncover "hidden" mechanism-based relations between diseases with no shared genes. For instance, two related diseases A and B may be associated with two separate sets of genes, $g(B)$ and $g(B)$, that are both part of a shared pathway. Therefore, similarity measures need to go beyond known gene associations. This could be done by either integrating other data in the analysis and/or by relating larger sets of genes to diseases. For example, Zhang and co-workers[7] used disease-gene association data from Genetic Association Database (GAD[8]) and protein-protein interaction (PPI) data from Human Protein Reference Database (HPRD[9]) to expand the HDN by adding new disease-gene associations inferred from the PPI. However, a better approach may be to make use of disease-pathway associations. A comparison between gene-based and pathway-based similarity measures can be found in the work of Lewis and colleagues[10] who used genome-wide association studies (GWAS[11]) to investigate disease similarity at different levels. They showed that pairwise similarities between diseases (measured by the Jaccard index, i.e. the number of shared genes/pathways divided by the total number of genes/pathways associated with the two diseases) are overall higher when pathways, rather than genes, are compared. Although pathway-based methods may be more helpful in identifying new disease-disease relations, in practice their applicability is limited because of the relatively low number of known disease-pathway associations. As proposed by Li and Agarwal[12], it is possible to infer such associations by using disease-gene and gene-pathway associations and by performing enrichment analyses to determine their statistical significance. However, because these inferences may not always be reliable, many investigators have used other methods to expand gene-disease associations.

Expression-based methods

A common way to link diseases with larger sets of genes is to use gene expression data obtained from Gene Expression Omnibus (GEO[13]). The main idea is to, for each disease, compare the diseased and normal samples to find the top differentially expressed genes. In a method proposed by Hu and Agarwal[14], the top genes with significant fold changes are determined for each disease, and enrichment analyses are performed to find significant overlaps between the sets of top genes identified with the diseases. Employing a similar approach, the web server DiseaseConnect[15] uses, in addition to gene expression profiles, GWAS[16] data and known gene-disease associations (from OMIM) to relate gene sets with diseases. Suthram *et al.*[17] developed a method in which fold changes of modules of functionally related genes are calculated by averaging those of the individual genes in the module. Each disease is then associated with a vector whose elements are the computed fold changes, and the similarity score of two diseases is defined as the partial Spearman correlation between their corresponding vectors. A similar algorithm, proposed by Yang and co-workers[18] and available as an R package called DSviaDRM[19], utilizes differential co-expression rather than expression.

Network-based methods

Although expression-based methods are capable of relating diseases that do not have known gene associations (a huge advantage over other methods), gene expression data are not available for most diseases. For this reason, these methods usually compute similarities between only tens or at most hundreds of diseases. Network-based approaches, on other hand, make larger scale studies of disease-disease associations possible. In a network-based method usually disease-gene associations are used to either connect the diseases to a molecular network (MN) or to associate each disease to a subnetwork of an MN. The MNs that so far have been used include functional gene-gene association networks (GGN) in which functionally related genes are linked, protein-protein interaction networks (PPN) where two proteins are connected if they are known to interact, and the human metabolic networks. One can build a disease-gene network (DGN) by connecting a disease A to its known associated genes $g(A)$ in a GGN. Similarly, a disease-protein network (DPN) may be constructed by linking each disease to the proteins encoded by its associated genes. In a DGN, it is possible to go beyond the disease genes (genes that are known to be associated with the diseases) and predict new disease-gene associations, potentially resulting in novel disease-disease relations. For example, Linghu and colleagues[20] utilized a Bayes classifier in conjunction with 16 genomic features to create a GGN with weighted links, from which the nearest genes to associated genes of each disease can be ranked and added to its corresponding list. The similarity between two diseases A and B is measured by how well the extended list of A, $G(A)$, captures the known associated genes of B, $g(B)$ and vice versa. Specifically, using $G(A)$ as an ordered hit list and $g(B)$ as the true positives, one obtains a Receiver Operating Characteristic curve, whose area under curve $AUC(A \to B)$ indicates the likelihood of predicting $g(B)$ from $G(A)$. The similarity is defined as $\sqrt{AUC(A \to B) \times AUC(B \to A)}$.

Lee *et al.*[21] chose the MN to be a human metabolic network, in which two compounds are linked if they are involved in a reaction. The network was constructed using the Kyoto Encyclopedia of Genes and Genomes (KEGG) Ligand database[22] and Biochemical, Genetic and Genomic

(BiGG) database of large scale metabolic reconstructions[23]. These databases list human metabolic reactions, the corresponding catalyzing enzymes, and their encoding genes. Lee *et al.* suggested to link two diseases if their associated enzymes (encoded by the associated genes) catalyze adjacent metabolic reactions (reactions involving a common metabolite). The resulting linked diseases were shown to have significantly higher comorbidity compared to other diseases.

Most network-based disease similarity measures use a PPN. PPNs have been used in conjunction with random walks to investigate disease-disease associations. In a method proposed by Hamaneh and Yu[2] random walkers start from each disease and walk on the DPN (built using disease-gene associations from the Comparative Toxicogenomics Database (CTD[24]) database and PPI from ppiTrim[25]) until they come back to the disease. The average number of visits to each protein is defined as its weight, and a vector is assigned to each disease whose components are these weights. The similarity between two diseases is then computed as the cosine of the angle between the corresponding vectors. Hamaneh and Yu showed that their algorithm, which has been implemented in the web service DeCoaD[26], is indeed capable of predicting novel disease-disease relations and proposed a new disease clustering algorithm. Suratanee and Plaimas[27] proposed to rank the proteins in a PPN (from the STRING[28] database) using random walks with restart, in which at each step random walkers (with a certain probability) may jump back to the starting points. For each disease A, random walks start from the associated proteins/genes (obtained from OMIM), and the ranks of all disease-associated proteins are determined according to average number of times they are visited. The similarity between diseases *A, B* is then calculated as $(1-(r_{AB}/N))\times(1-(r_{BA}/N))$ where r_{AB} is the median rank of all proteins associated with *B* relative to *A,* and *N* is the total number of disease proteins.

Some studies use PPN to define a topological similarity/distance measure based only on the known gene/protein associations instead of expanding these associations. Sun and co-workers[29], for example, suggested a rather complicated node similarity measure to compute pairwise similarity scores between the proteins associated with different diseases, which can be combined to calculate the disease-disease similarities. The authors reported a comparable performance compared to the simple method based on gene sharing, although they argued that the two methods provide complementary information. In a recent study[30] an extended PPN (including binary protein-protein, regulatory, metabolic, and other interactions) has been utilized to define the *distance* between two protein lists associated with two diseases that have large (> 20) number of gene associations. The authors argued that this condition is essential, because when the number of associated proteins is large they form one or more connected subgraphs in the PPN. Therefore, for disease A a diameter, denoted byd_A, can be defined as the average shortest distance (on the PPN) between the proteins associated with the disease. Subsequently, the distance s_{AB} between two diseases A and B can be defined as $\mathrm{s_{AB}} = \mathrm{d_{AB}} - (1/2)(\mathrm{d_A} + \mathrm{d_B})$, where d_{AB} is the average shortest distance between the members of the two protein sets. The authors showed that their distance measure agrees well with a few available biological data (comorbidity, for example). Although the results of this study provide insight into the relation between disease-disease distances and protein interactions, the method is not useful for finding diseases similar to newly discovered or rare illnesses that have very few gene associations.

Methods that go beyond disease-gene associations

The methods discussed so far (except the ones based on gene expressions) require known disease-gene associations, i.e. they cannot be applied to diseases without such associations. On the other hand, as mentioned before, the applicability of expression-based similarity measures are limited. Therefore, methods that are not expression-based and do not require known gene associations may be very helpful. However, there are very few such methods, and they still use disease-gene associations as an optional input. Kim and colleagues[31] employed a text-mining approach to find disease-gene and disease-drug associations by counting the number of co-appearances (frequencies) of disease/gene or disease/drug pairs in the literature. Their method assigns two vectors to each disease whose components are the aforementioned frequencies. For each disease pair two similarity measures (one gene-based and one drug-based) are then computed based on the mutual information of their corresponding vectors. The two similarity scores are then combined to determine the total score. Although this method can be applied to a large number of diseases, it was evaluated using a very small subset of disease pairs. Therefore the efficacy of the method (especially for diseases with no known gene associations) remains to be tested. In another study of this type, Sun *et al.*[32] used disease-chemical in addition to disease-gene associations for elucidating disease-disease associations. They also included disease-pathway and disease-GO (Gene Ontology[33]) term associations inferred from disease-gene, gene-pathway, and gene-GO term relations. The proposed model assigns a vector to each disease representing its associations, and disease-disease similarities are calculated based on the cosine of the angles between the corresponding vectors. The authors evaluated their method by comparing their results with co-occurrence of disease names in the literature. They showed that integrating all associations together results in a better performance comparing to the case only one type of data is considered. Interestingly, they also reported a slightly better performance for the chemical-based (using only disease-chemical associations) similarity compared to the gene-based similarity.

Our survey of mechanism-based disease similarity measures implies that the vast majority of methods can be summarized in the same way: the similarity between two diseases is calculated by comparing sets of genes (or their corresponding proteins) associated with the diseases. The differences between these approaches are only in the way the genes are associated to the diseases or in how the gene sets are compared. Therefore, as the works presented in 31 and 32 suggest, integrating other data in the process may be helpful to reveal new disease-disease relations, especially in the case of rare diseases that usually have few gene associations.

Acknowledgements

This work was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health.

References

1. Zhou X, Menche J, Barabási AL, Sharma A. Human symptoms-disease network. Nat Commun. 2014; 5:4212-4221.
2. Hamaneh MB, Yu YK. Rleating diseases by integrating gene associations and information flow through protein interaction network. PLoS One. 2014; 9:e110936.
3. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proc Natl Acad Sci USA. 2007; 104:8685-8690.
4. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005; 33:D514-517.
5. Bauer-Mehren A, Bundschus M, Rautschka M, Mayer MA, Sanz F, Furlong LI. Furlong. Gene-disease network analysis reveals functional modules in Mendelian, complex and enviornmental diseases. PLoS One. 2011; 6:e20284.
6. Park J, Lee DS, Christakis NA, Barabási AL. The impact of celllular networks on disease comorbidity. Mol Syst Biol. 2009; 5:262-268.
7. Zhang X, Zhang R, Jiang Y, Sun P, Tang G, Wang X. The expanded human disease network combining protein-protein interaction information. Eur J Hum Genet. 2011; 19:783-788.
8. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nat Genet. 2004; 36:431-432.
9. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, *et al*. Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res. 2004; 32:D497-D501.
10. Lewis SN, Nsoesie E, Weeks C, Qiao D, Zhang L. Prediction of disease and phenotype associations from genome-wide association studies. PLoS One. 2011; 6:e27175.
11. Johnson AD, O'Donnell CJ. An open access database of genome-wide association results. BMC Med Genet. 2009; 10:1-6.
12. Li Y, Agarwal P. A pathway-based view of human diseases and disease relationships. PLoS One. 2009; 4:e4346.
13. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: mining tens of millions of expression profiles—database and tools update. Nucleic Acids Res. 2007; 35:D760-D765.
14. Hu G, Agarwal P. Human disease-drug network based on genomic expression profiles. PLoS One. 2009; 4:e6536.
15. Liu CC, Tseng YT, Li W3, Wu CY, Mayzus I, Rzhetsky A, et al. DiseaseConnect: a comprehensive web server for mechanism based disease-disease coonections. Nucleic Acids Res. 2014; 42:W137-W146.
16. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014; 42:D1001–D1006.
17. Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, Butte AJ. Network-based elucidation of human disease similarities reveals functional modules enriched for pluripotent drug targets. PLoS Comput Biol. 2010; 6:e1000662.
18. Yang J, Wu SJ, Dai WT, Li YX, Li YY. The human disease network in terms of dysfunctional regulatory mechanisms. Biol Direct. 2015; 10:60-81.
19. Yang J, Wu SJ, Li YX, Li YY. DSviaDRM: an R package for estimating disease similarity via dysfunctional regulation mechanism. Bioinformatics. 2015; 31:3870-3872.
20. B. Linghu, E. Snitkin, Z. Hu, Y. Xia and C. Delisi. Genome-wide prioritization of disease genes and identification of disease-disease association from an integrated human functional linkage network. Genome Biol.2009; 10:R91-R117.
21. Lee DS, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabási AL. The implications of human metabolic network topology for disease comorbidity. Proc Natl Acad Sci USA. 2008; 105:9880-9885.
22. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 2006; 34:D354-D357.
23. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA. 2007; 104:1777-1782.
24. Davis AP, Grondin CJ, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, *et al*. The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. Nucleic Acids Res. 2015; 43:D914-D920.
25. Stojmirović A, Yu YK. ppiTrim: constructing non-redundant and up-to-date interactomes. Database. 2011; 2011:bar036.
26. Hamaneh MB, Yu YK. DeCoaD: determining correlations among diseases using protein interaction networks. BMC Res Notes. 2015; 8:226:232.
27. Suratanee A, Plaimas K. DDA: A novel network-based scoring method to identify disease-disease associations. Bioinform Biol Insights. 2015; 9:175:186.
28. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, *et al*. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 2011; 39:D561-D568.
29. Sun K, Gonçalves JP, Larminie C, Przulj N. Predicting disease associations via biological network analysis. BMC Bioinfomatics. 2014; 15:304-316.
30. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015; 347:1257601.
31. Kim H, Yoon Y, Ahn J, Park S. A literature-driven method to calculate similarities among diseases. Comput Methods Programs Biomed. 2015; 122:108-122.

32 Sun K, Buchan N, Larminie C, Pržulj N. The integrated disease network. Integr Biol. 2014; 6:1069-1079.

33. Ashburner M1, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25:25-29.