Constructing and analyzing a disease network based on proteins

Ying Li^{1,a}, Wencong Huang²

¹Guangxi Colleges and Universities Key Laboratory of Scientific Computing & Intelligent Information Processing, Nanning Normal University, Nanning 530299, China

²School of Physical Education and Health, Nanning Normal University, Nanning 530299, China

Abstract. Protein is the specific executor of life activities, but there is no protein-based disease network and the current disease networks cannot show that a disease group share the same factors. We propose a method to construct a protein-based network by assigning disease pairs to different intervals according to their similarities and searching for disease groups in each interval. Statistical methods are used to analyze the disease network, and the result indicates that : in the case where a disease belongs to only one disease group, most diseases have their own protein characteristics, but the common protein of them is not obvious; the more diseases a protein is related to, the more likely the protein becomes common protein; diseases grouping at protein level in this study are different from traditional disease classification; there is a certain relationship between disease symptoms and underlying proteins, but not one-to-one correspondence.

1 Introduction

Traditional classification of human diseases is based on the observational correlation between the existing knowledge of pathological analysis and clinical symptoms. Its drawbacks are lack of sensitivity to identify latent diseases, clear definition of disease specificity, and neglect of the correlation of diseases. A system-based network framework was proposed to correct these shortcomings. It provides a new framework for classifying diseases, defining diseases clearly, predicting disease outcomes and determining tailored treatment strategies [1]. The human disease network is a subnet of this network framework. There are two types of disease networks. One is a network with diseases and related factors as nodes, reflecting the relationship between diseases and multiple pathogenic factors. Such disease networks include the cryptorchidism syndrome network [2], the retinal disease network [3], the chronic obstructive pulmonary disease network [4] and ulcerative colitis disease network [5], etc. Another network is using a disease as a node, in which two diseases are linked by their shared cellular components, biological processes or other disease-related factors, reflecting the relationship between diseases and diseases. We study the second type of disease network (known as Diseasome). Diseases may share cellular components such as genes, proteins, microRNAs, biological processes such as metabolism, and disease-related factors such as phenotypes, symptoms and environmental factors.

Currently, the human disease networks based on single biological related factors include: the genetic human disease network [6,7,8], the metabolic disease network [9], the microRNA-related disease

network[10,11], the phenotypic disease network [12,13,14], the symptom disease network [15], the differential co-expression-based disease network [16,17], the environment-based disease network[18], the human pathway disease network [19], etc. The networks based on multiple biological related factors include: the disease network based on genetic and environmental factors [20], and the disease network constructed by combining genetic, morbidity and co-morbidity information from clinical outcomes, and metabolic pathways [21], etc. There is a special disease network constructed in which two diseases are linked if their observed probability of co-occurrence deviating that expected under independence [22].

Because of the existence of gene regulation mechanism, gene expression and protein level are not entirely consistent. Protein is the specific executor of life activities. However, there is no disease network based on protein. Constructing such disease network is helpful to understand that different phenotypes of appellations in different medical branches are linked at the protein level, and why certain diseases occur in a group. The co-morbidity conditions of diseases in the same group provide insights into possible new methods for disease prevention, diagnosis and treatment, especially for drug redirection.

The disease networks can be classified to two types according to the connection mode among their nodes. One is to simply link diseases sharing a certain factor without calculating the similarity between two diseases; the other is to link diseases when their similarity exceeds a given threshold. The networks constructed in the first way include the human disease network sharing genes, the metabolic disease network, the microRNA-related disease network, the disease network based on

^a Corresponding author: 2313166529@qq.com

genetic and environmental factors, and the microRNA-gene-disease network, etc. The networks constructed in the second way include the phenotypic disease network, the symptomatic disease network and the disease network based on differential co-expression, etc. The first way is simple and easy to implement, but it cannot distinguish the strength of the link between diseases; the second way tends to reflect the close relationship between diseases, in which the choice of threshold is the difficulty and key point. The disease networks constructed in both ways may contain the aggregation of multiple diseases, but these diseases do not necessarily share the same factors. This is because the construction of these disease networks is based on the similarity between two diseases, without further exploring the common factors shared by multiple diseases.

To ensure that diseases clustered in disease networks share the same factors, we propose a method in which disease similarities are assigned to different intervals and then search for disease groups in each interval to construct a protein-based disease network. Then we analyze the constructed disease network. All diseases in a group in our study are related to the same proteins. A disease group forms a complete subgraph in the protein-based disease network. The disease network constructed by searching disease groups can accurately and intuitively show that multiple diseases share the same proteins. Furthermore, searching for disease groups in intervals can also reduce the search space for looking for diseases, thus improving the speed of constructing disease network.

2 Materials and Methods

2.1 Data source and basic processing

Papers with titles of proteins, enzymes, transcription factors, receptors and antibodies were obtained in PubMed literature library. The abstracts of these papers were submitted to SemReq software to extract the connections between content words. A total of 1665 2122 connections were obtained. According to the disease list obtained from Mesh disease catalog and the words of protein, enzyme, transcription factor, receptor and antibody, 8151 effective connections of diseases and proteins were extracted from the above mentioned connections (words containing non-disease or non-protein have been filtered out), including 963 proteins and 670 diseases (diseases not in MeSH disease list have been filtered out). If two diseases are connected with the same protein, link them together. In this way, a total of 17994 disease-disease links are obtained from the disease-protein linkage records.

2.2 Computation of disease similarity

Diseases can be expressed as vectors $D_i = \{d_{ij} | j=1, 2, \dots, n\}$, d_{ij} is 0 or 1, d_{ij} is 1 when a disease is related to protein i , and 0 when it does not. The cosine similarity formula is used to calculate the similarity between two diseases. Let the vectors of two diseases to be D_1 and D_2 , $D_1 = (d_{11},$

$d_{12}, \dots, d_{1m})$ and $D_2 = (d_{21}, d_{22}, \dots, d_{2m})$. The cosine similarity formula of the two diseases is the following.

$$\cos \theta = \frac{\sum_{i=1}^m d_{1i} \times d_{2i}}{\sqrt{\sum_{i=1}^m d_{1i}^2} \times \sqrt{\sum_{i=1}^m d_{2i}^2}} \quad (1)$$

Since the values of d_{1i} and d_{2i} are 0 or 1, the formula is transformed into the following one.

$$\cos \theta = \frac{|\{d_{11}, d_{12}, \dots, d_{1m}\} \cap \{d_{21}, d_{22}, \dots, d_{2m}\}|}{\sqrt{|\{d_{11}, d_{12}, \dots, d_{1m}\}|} \sqrt{|\{d_{21}, d_{22}, \dots, d_{2m}\}|}} \quad (2)$$

Numerator in the formula represents the number of proteins which both disease D_1 and disease D_2 are related to. Square root symbol on the left in the denominator represents the number of proteins which disease D_1 are related to, and that in the right represents the number of proteins which disease D_2 are related to. The similarities of 17994 disease-disease links obtained in Section 2.1 are calculated using formula (2).

2.3 The method to construct the protein-based disease network

When searching for a disease group in an interval, using the principle of maximum similarity priority to select two diseases with the maximum similarity as the initial two diseases of a disease group, and the proteins being related to the two diseases are used as the initial proteins of the common protein set of the disease group. The candidate diseases in the interval are investigated one by one. If a candidate disease satisfies two conditions: one is that the disease is linked to a disease in the current disease group; the other is that after the disease joins the current disease group, the updated set of common proteins cannot be empty. The two conditions ensure that the candidate disease and the current disease groups form a complete graph in which diseases share the same proteins. The investigated disease is added to the current disease group, and the common protein set is updated with the intersection of the disease-related protein set of the added disease and the current common protein set, and the investigated disease is deleted from the candidate disease set. The candidate diseases are investigated until no eligible disease can be found, then the next disease group are looked for. When the disease pairs with the maximum similarity can no longer be found in the current interval, stop searching for disease groups. Each disease in the current candidate disease set of the interval becomes a disease group in which the common proteins is the related proteins of itself. All disease complete subgraphs required from all intervals constitute the whole disease network.

3 Results

3.1. Generating a disease network based on proteins

There is 670 diseases from the connections between diseases and proteins, but there is 626 diseases from 17994 disease pairs, indicates that 44 diseases had no connections with other diseases. After searching for disease groups in 10 intervals, 235 disease groups with at least two diseases and 49 disease groups with only one disease are required. Therefore, we get 93 disease groups with only one disease. In this study , each disease belongs to only one disease group.

The whole disease network is shown in figure 1, and each complete subgraph in the network is a disease group. The circle surrounded by nodes in the graph represents a disease similarity interval. Seven intervals contain more disease subgraphs than the remainder. The common feature of the seven intervals is that the number of complete subgraph containing two diseases is the largest, and the number of subgraph containing more than three diseases is relatively small. Interval (0.0, 0.1] contains only one subgraph in it and only two diseases in

the subgraph. Interval (0.6, 0.7] had three disease groups and interval (0.8, 0.9] had six disease groups. The number of diseases in these complete subgraphs of the above three intervals is 2 or 3. The complete subgraph containing the largest number of diseases is in interval (0.9, 1.0]. Similarity of any two diseases in the interval is 1, so it is the maximum similarity interval. The interval has one subgraph of 10 diseases, two subgraphs of 7 diseases and one subgraph of 6 diseases. The interval also has disease subgraphs with 5, 4, 3 and 2 diseases, and the number of subgraphs with 2 diseases is the largest. There are a considerable number of isolated nodes in the whole disease network, which are groups of diseases formed by a single disease. In the whole network, about 94% of the disease groups only contain 1-3 diseases, and only about 6% of the disease groups contain more than 3 diseases. This indicates that a small number of disease groups contain much more diseases, while most disease groups contain a few diseases.

Figure 1.The disease network based on proteins. Different color nodes are used to represent different disease similarities. Each white node represents a disease which is a group. The enlarged image of the light blue region consists of three subnets, the disease within the three disease groups and the common proteins of them are shown in the list in the right.

3.2 Relationship between the number of common proteins and the number of disease groups

To reveal the relationship between the number of common proteins and the number of disease groups, add up the number of disease groups group by the number of

common protein of them. Do this on 10 intervals respectively, and do on (0.0,1.0] last. The results are shown in pane a to panel k of figure 2. Eight of pane a-j shows a decreasing trend in the number of disease groups with the increase of the number of common proteins; only panel b and d do not conform to this rule. Panel k shows that the number of disease groups decreases with the increase of the number of common

proteins. There are about 71.5% of the total disease group with 1 common protein; about 16.6% disease groups with 2 common proteins; when the number of common protein reaches 26 and 29, the number of disease groups is 1 respectively. The disease group with 29 common proteins contains "Carcinoma, Non-Small-Cell Lung" and "Glioblastoma". The former is related to

108 proteins, while the latter is related to 72 proteins. The results show that there are a large number of disease groups with a small amount of common proteins, but a small number of disease groups with a large amount of common proteins. This indicates that most diseases have their own protein characteristics, but the common proteins of them is not obvious.

Figure 2. Relationship between the number of common proteins and the number of disease groups.

3.3 Relationship between the number of diseases a protein being related to and the number of disease groups it being their common protein

In order to reveal the relationship between the number of diseases a protein being related to and the number of disease groups it being their common protein, add up the number of diseases which proteins are related to group by proteins, and the result is arranged in descending order. Then calculate the number of disease groups for each protein if the protein has been a common protein. Statistical results show that there are 963 proteins which are related to diseases, and only 219 common proteins (about 22.8% of the 963 proteins), indicating that only a small number of proteins become common proteins. In figure 3 A, as the number of diseases decreases, so does the number of disease groups. This indicates that the more diseases a protein is related to, the more disease groups of which the protein is common protein, and vice versa. Calculate the Pearson correlation coefficient of the number of diseases and the number of disease groups associated with the common proteins (219 proteins), and $p \approx 0.912$ is obtained. This suggests that the two are positively correlated, that is, the more diseases a proteins is related to, the more disease groups of which the protein is common protein. B and C in figure 3 show that 14 of the top 20 proteins with the largest number of associated diseases appear in the top 20 common

proteins with the largest number of associated disease groups.

3.4 Relationship between disease groups and MeSH disease categories

In MeSH disease catalog, a disease can belong to more than one disease categories, while a disease can only belong to a disease group in this study. In order to reveal the relationship between disease groups and MeSH disease categories, we analyze their mutual inclusion. Firstly, for each disease group which contains at least two diseases (235 groups in total), calculate how many diseases in a group belonging to the same MeSH category. The result is shown in figure 4 A. There is 69 disease groups for which all diseases in each group belong to the same MeSH disease category (about 29% of the total number of disease groups). For example, in the disease groups with similarity 1.0, there is a group of three diseases including "Paraparesis", "Spastic Spina Bifida Occulta", and "Epilepsy, Rolandic". All these three disease belong to the "Nervous System Diseases (C10)" of MeSH, and their common protein is "GABA Receptor" (gamma-aminobutyric acid receptor). This protein is an important inhibitory neurotransmitter, which participates in various metabolic activities and has high physiological activity. There is 27 disease groups for which half and more but not all diseases in each group belong to the same MeSH disease category (about 12% of the total number of disease groups). There is 139 disease groups for which less than half diseases in each

group belong to the same MeSH disease category (about 59% of the total number of disease groups). For example, the densest disease groups with similarity of 1 consist of 10 diseases, which belong to 12 MeSH disease categories. This indicates that most diseases in a disease

group belong to more than one MeSH disease categories. Grouping diseases at protein level in this study are different from traditional disease classification to a large extent.

Figure 3. Relationship between the number of disease a protein being related to and the number of disease groups it being their common protein. The blue curve in A is about the number of diseases which proteins are connected with, and the green curve is about the number of disease groups of which proteins are common proteins.

Secondly, calculate each MeSH disease category containing how many disease groups for which diseases in a group belong to the same MeSH disease category (69 disease groups in all) to understand which disease group belong to which MeSH category. Results are shown in figure 4 B. 21 of the 26 MeSH disease categories contain some of the disease groups. The top three MeSH disease categories with the largest number disease groups are C04 (Neoplasms), C10 (Nervous System Diseases) and C23 (Pathological Conditions, Signs and Symptoms), among which C04 contains 15

disease groups, and C10 contains 11 disease groups and C23 contains 9 disease groups. This suggests that there are more or less diseases in most MeSH disease categories having the same classification on protein level with traditional disease classification. There are more diseases in C01, C10 and C23 having the same classification on protein level with traditional disease classification compared with other MeSH disease categories.

Figure 4. Relationship between disease groups and MeSH disease categories.

3.5 Relationship between symptoms and proteins

Some diseases in a disease group have the same symptoms, but some diseases in the same group have different symptoms. For example, in the densest disease group which contains 10 diseases, some diseases have the same inflammatory symptom, such as "Lymphocytic Choriomeningitis", "Candidiasis, Cutaneous" and "Molluscum Contagiosum". Some have different symptoms, such as "hypersplenism" and the three diseases mentioned above. Symptoms of "hypersplenism" is splenomegaly. If it is congestive splenomegaly, there is no inflammation. This suggests that "Hypersplenism" and the above three diseases have different symptoms. The 10 diseases are only related to the protein "TLR4 protein, human". This example indicates that the protein does not correspond to a specific symptom. This phenomenon is related to the conditions under which the protein is activated. Inflammation is often accompanied by exogenous pathogens, while inflammation is not necessarily caused by endogenous substances.

Diseases from different disease groups share the same symptoms, but underlying proteins which they are associated with are different. For example, "Cholera" and "Rotavirus Infections" are both symptoms of diarrhea, but the protein being associated with Cholera is "GTP-Binding Proteins", while the protein being associated with "Rotavirus Infections" is "Membrane Glycoproteins". This example shows that the symptoms of the two diseases are the same, but the underlying proteins which they are associated with are not necessarily the same.

The above three examples show that there is a certain relationship between disease symptoms and underlying proteins which these diseases are related to, but not one-

to-one correspondence. Some diseases in a disease group share the same symptoms, while some diseases in the same group do not share the same symptoms. Diseases from different disease groups share the same symptoms, but may be associated with different proteins.

4 Conclusion

In this study, we found that most diseases have fewer protein commonalities and more protein characteristics. There are a lot of disease groups with a small amount of common protein, and a small number of disease groups with a large amount of common protein, which means that the protein characteristics of most diseases are more prominent and the protein commonalities are not obvious. Statistics show that diseases associated with more proteins are clustered with a few diseases. It indicates that the more proteins a disease is related to, the more obvious its protein characteristics. It also suggests that the pathogenic factors of these diseases are complex. On the contrary, those diseases being related to less proteins are more likely to cluster with more diseases, indicating that the protein characteristics of those diseases are not obvious and suggesting that the pathogenic factors of them are relatively simple. The number of dense disease groups is very small, and the number of sparse disease groups is very large, which also suggests that the protein commonness among most diseases is less, and their protein characteristics are more, so only a few disease groups can be clustered into a few dense groups, while more diseases can only form many sparse disease groups.

The more diseases a protein is related to, the more likely it becomes a common protein. By calculating Pearson correlation coefficient between the number of diseases a protein being related to and the number of disease groups of which the protein are common protein, we found that the more diseases a protein is related to, the more disease groups it is their common protein and

vice versa. In other words, the more diseases a protein is connected with, the more likely it is a common protein.

There is still room for exploration classification methods of diseases. In this study, for a small number of disease groups, the diseases in each group belong to the same of MeSH disease category, which indicates that a few diseases have the same classification at the level of protein as that of traditional disease classification. For most disease groups, diseases in each group belong to multiple MeSH disease categories, which indicates that the most diseases have different classification at the level of protein as that of traditional disease classification. There is still room for exploration on how to classify diseases.

The relationship between disease symptoms and underlying proteins is complex and not one-to-one. After preliminary study, we found that there is a certain correlation between the two, but not one-to-one correspondence, indicating that the underlying proteins which some diseases are related to are the same, but the symptoms of these diseases are not necessarily the same; while the symptoms of some diseases are the same, the underlying proteins which they are related to are not necessarily the same. Although it is not yet clear what the association is, the results of this study provide clues for further research.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 61741203, Grant No. 61866006); the Natural Science Foundation of Guangxi province of China (Grant No. 2016GXNSFAA380243).

References

- Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet*, 12(1):56-68(2011).
- Urh K, Kolenc Ž, Hrovat M. et al. Molecular Mechanisms of Syndromic Cryptorchidism: Data Synthesis of 50 Studies and Visualization of Gene-Disease Network. *Front Endocrinol (Lausanne)*, 9, doi: 10.3389/fendo.2018.00425(2018).
- Lázaro-Guevara JM, Flores-Robles BJ, Garrido K. et al. Gene's hubs in retinal diseases: A retinal disease network. *Heliyon*, 6, e00867(2018).
- Sharma A, Kitsak M, Cho MH. et al. (2018) Integration of Molecular Interactome and Targeted Interaction Analysis to Identify a COPD Disease Network Module. *Sci Rep*, 8, doi: 10.1038/s41598-018-32173-z(2018).
- Jodeleit H, Palamides P, Beigel F. et al. Design and validation of a disease network of inflammatory processes in the NSG-UC mouse model. *J Transl Med.*, 15, doi: 10.1186/s12967-017-1368-4(2017).
- Goh KI, Cusick ME, Valle D. et al. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21): 8685-8690(2007).
- Suk K. Human-yeast genetic interaction for disease network: systematic discovery of multiple drug targets. *BMB Rep.* 50(11):535-536(2017).
- Lo Surdo P, Calderone A, Iannuccelli M. et al. DISNOR: a disease network open resource. *Nucleic Acids Res.*, 46(D1):D527-D534. doi: 10.1093/nar/gkx876(2018).
- Lee DS, Park J, Kay KA. et al. The implications of human metabolic network topology for disease comorbidity. *PNAS*, 105:9880–9885(2008).
- Lu M, Zhang Q, Deng M. et al. An Analysis of Human MicroRNA and Disease Associations. *Plos ONE*, 3, e3420(2008).
- Yuan D, Cui X, Wang Y. et al. Enrichment Analysis Identifies Functional MicroRNA-Disease Associations in Humans. *PLoS ONE*, 10, e0136285. doi: 10.1371/journal.pone.0136285(2015).
- Van Driel MA, Bruggeman J, Vriend G. et al. A text-mining analysis of the human phenotype. *European Journal of Human Genetics*, 14:535–542(2006).
- Rzhetsky A, Wajngurt D, Park N. et al. Probing genetic overlap among complex human phenotypes. *PNAS*, 104:11694–11699(2007).
- Hidalgo CA, Blumm N, Barabási AL. et al. A Dynamic Network Approach for the Study of Human Phenotypes. *Plos Computational Biology*, 5, e1000353(2009).
- Zhou X, Menche J, Barabási AL. et al. Human symptoms-disease network. *Nat. Commun.*, 5:4212, doi: 10.1038/ncomms5212(2014).
- Yang J, Wu SJ, Dai WT. et al. The human disease network in terms of dysfunctional regulatory mechanisms. *Biol Direct*, 10:60. doi: 10.1186/s13062-015-0088-z(2015).
- Yang J, Wu SJ, Yang SY. et al. DNetDB: The human disease network database based on dysfunctional regulation mechanism. *BMC Syst Biol*, 10(1):36. doi: 10.1186/s12918-016-0280-5(2016).
- Taboureau O, Audouze K. Human Environmental Disease Network: A computational model to assess toxicology of contaminants. *ALTEX*. 34(2):289-300. doi: 10.14573/altex(2017).
- Yu L, Gao L. Human pathway-based disease network. *IEEE/ACM Trans Comput Biol Bioinform*, 17, 2019, doi: 10.1109/TCBB.2017.2774802(2017).
- identification and clustering of human disease etiological factors. *BMC Bioinformatics*, 10 Suppl 2:S14. doi: 10.1186/1471-2105-10-S2-S14(2009).
- Bang S, Kim JH, Shin H. Causality modeling for directed disease network. *Bioinformatics*, 32(17):i437-i444. doi: 10.1093/bioinformatics/btw439(2016).
- Jiang Y, Ma S, Shia BC. et al. An Epidemiological Human Disease Network Derived from Disease Co-occurrence in Taiwan. *Sci Rep.*, 8(1):4557, doi: 10.1038/s41598-018-21779-y(2018).