BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 11(9)

Modular organization of the human disease genes: a text-based network inference

Hongdong Xuan[1,2], Xin Li[2], Shenrong Ren[3], Shihua Zhang[1]*

[1]Department of Biostatistics, School of Science, Anhui Agricultural University, Hefei 230036, China; [2]College of Information and Computer science, Anhui Agricultural University, Hefei 230036, China; [3]School of life sciences, Anhui Agricultural University, Hefei 230036, China; Shihua Zhang - Email: zhangshihua@ahau.edu.cn; Fax: +86 551 6578 6121; *Corresponding author

Received July 29, 2015; Revised September 06, 2015; Accepted September 09, 2015; Published September 30, 2015

Abstract:
The analysis of disease phenotype data with genetic information indicated that genes associated with clinically similar diseases tend to be functionally related and work together to perform a specific biological function. Therefore, it is of interest to relate disease phenotype data to mirror modular property implied in the association map of disease genes. Hence, we constructed a text-based human disease gene network (HDGN) by using the phenotypic similarity of their associated disease phenotype records in the OMIM database. Analysis shows that the network is highly modular and it is highly correlated with the physiological classification of genetic diseases. Using a graph clustering algorithm, we found 139 gene modules in the network of 1,865 genes and their gene products (proteins) in these gene modules tend to interact with each other via the computation of PPI intensity. Genes in such gene modules are functionally related and may represent the shared genetic basis of their corresponding diseases. These genes, alone or in combination, could be considered as potential therapeutic targets in future clinical therapy.

Key words: disease phenotype, text-mining, modularity, genetic basis.

Background:

When used together with genetic information, phenotype data can help to explore relationships between genetic diseases and mutation-bearing genes **[1].** However, phenotype data such as Online Mendelian Inheretance in Man (OMIM) **[2]** and PhenomicDB **[3]** remain intractable to be deal with because the lack of a standardized vocabulary for the phenotype description. Despite these difficulties, there exists some successful groundwork in utilizing such daunting phenotype data. For example, Freudenberg *et al.* **[4]** clustered nearly 1,000 disease phenotypes of known genetic origin from OMIM according to their phenotypic similarity using periodicity, etiology, tissue, age of onset and mode of inheritance as classification indices. Their results showed that genes causing similar disease phenotypes have similar Gene Ontology (GO) functional annotation. Groth *et al.* **[5]** used text clustering to group genes based on their phenotype data from PhenomicDB. The results indicated that these clusters correlate with several indicators for biological coherence in gene groups, such as GO functional annotation and protein-protein interaction (PPI). Both the investigations revealed a fact that genes associated with similar disease phenotypes are more likely to be functionally related. These related genes work together, as a functional module, such as protein complex and cell pathway, to perform a specific biological function **[6].** The functional relationship in these related genes are in agreement with the modular property of most biological networks, indicating the existence of densely-connected subgraphs in the gene functional network.

In this study, we constructed a functional network of human disease genes, and further investigated its modular property. We determined the association between disease genes in the network by using their phenotypic relatedness. The human disease gene network (HDGN) has been proved to have a high modular architecture. From the network, we extracted 139 gene modules and found the modularity correlates with the functional level of PPI. Of these 139 gene modules, 127 (91.4%)

were significantly enriched in only one disease class or two. Therefore, our network-based framework revealed that disease genes and the associated genetic diseases have a high level of agreement in functional interplay, although they are at two different biological levels.

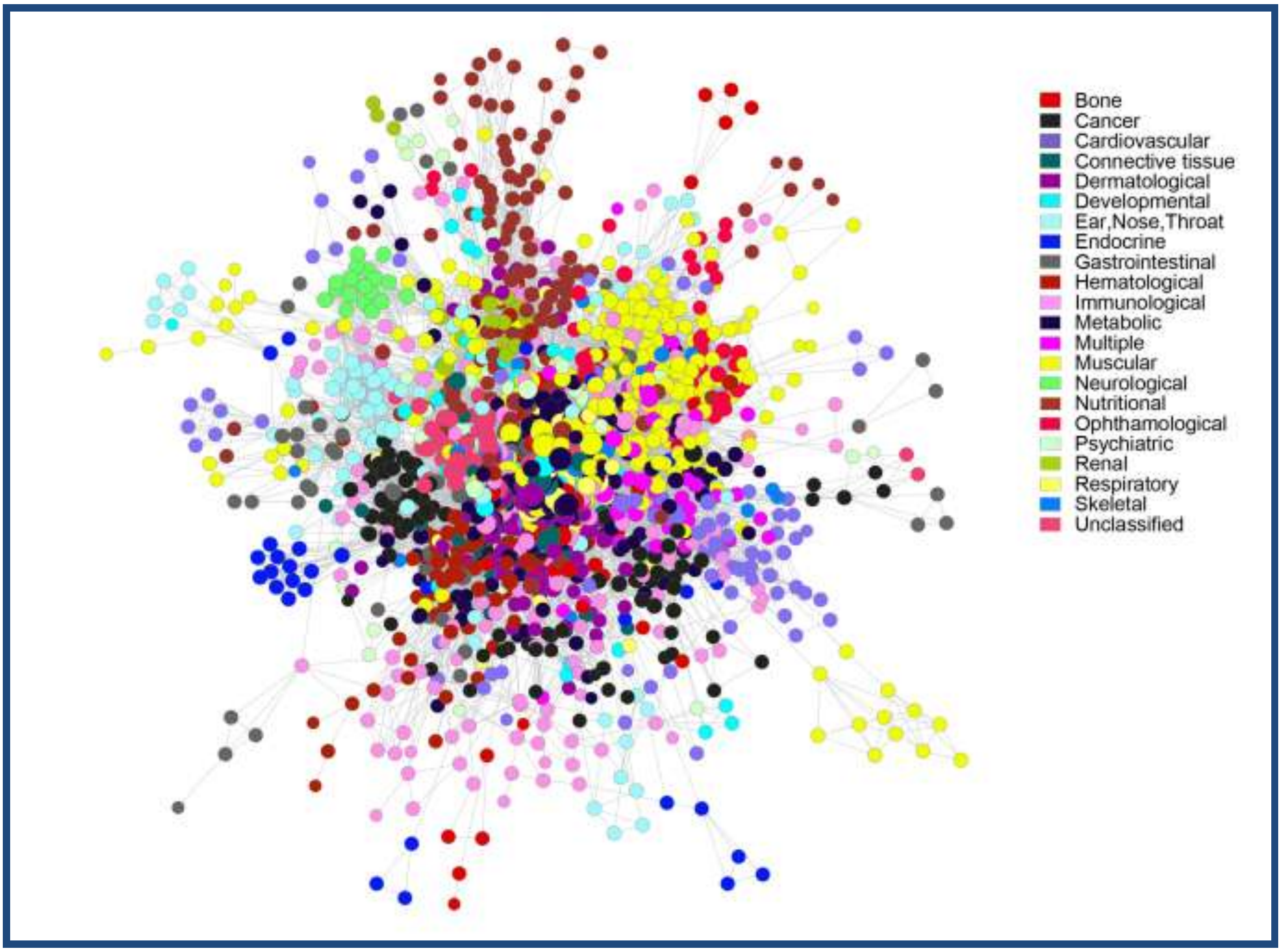

Figure 1: The giant connected component of HDGN in the HDGN, the color of a disease gene node depends on the disease class to which the associated disease phenotype belongs. There are 22 disease classes for which their names, assigned colors are shown on the right of the Figure. It is noted that we referred to the disease classification described by Gol et al. (13) who manually classified disease phenotypes according to the physiological system affected.

Methodology:

Disease phenotype similarity measure

In OMIM, we considered the combination of text (TX) and clinical synopsis (CS) fields as a full phenotype record. Phenotype records were parsed by the MetaMap Transfer tool **[7]**, a configurable program to map text to the Unified Medical Language System (UMLS) Metathesaurus concepts **[8]**. Thus, phenotype records could be referred to as phenotype feature vectors. In this work, we used the term frequency–inverse document frequency weighting scheme **[9]** for the refinement of phenotype feature vectors and the cosine similarity measure for calculating the phenotypic similarity between different phenotype records.

Construction of HDGN

Inspired by the fact that genes associated with similar disease phenotypes are likely to be functionally related, we used the phenotypic relatedness to decide the functional relatedness of disease genes. The association between two disease genes in HDGN was decided when the phenotypic similarity score of their associated disease phenotypes exceeded the significant cutoff. The cutoff was chosen based on the random shuffling of the phenotype feature vectors of the two disease phenotypes and similarity score ranking.

Modular measure

We used two modular measures, dyadicity *D* and heterophilicity *H*, whch were proposed by Park *et al.* **[10]**, to quantify the modular property of HDGN. Dyadicity is a measure of the enrichment of links between nodes sharing a common property over the number expected if the characteristics were distributed randomly on the network. Heterophilicity is a measure of the tendency of nodes to connect with other nodes with a common property. In HDGN, disease genes with their associated disease phenotypes belonging to the same disease class were regarded to have the common property. Thus, we can compute the *D*s and *H*s for different disease classes.

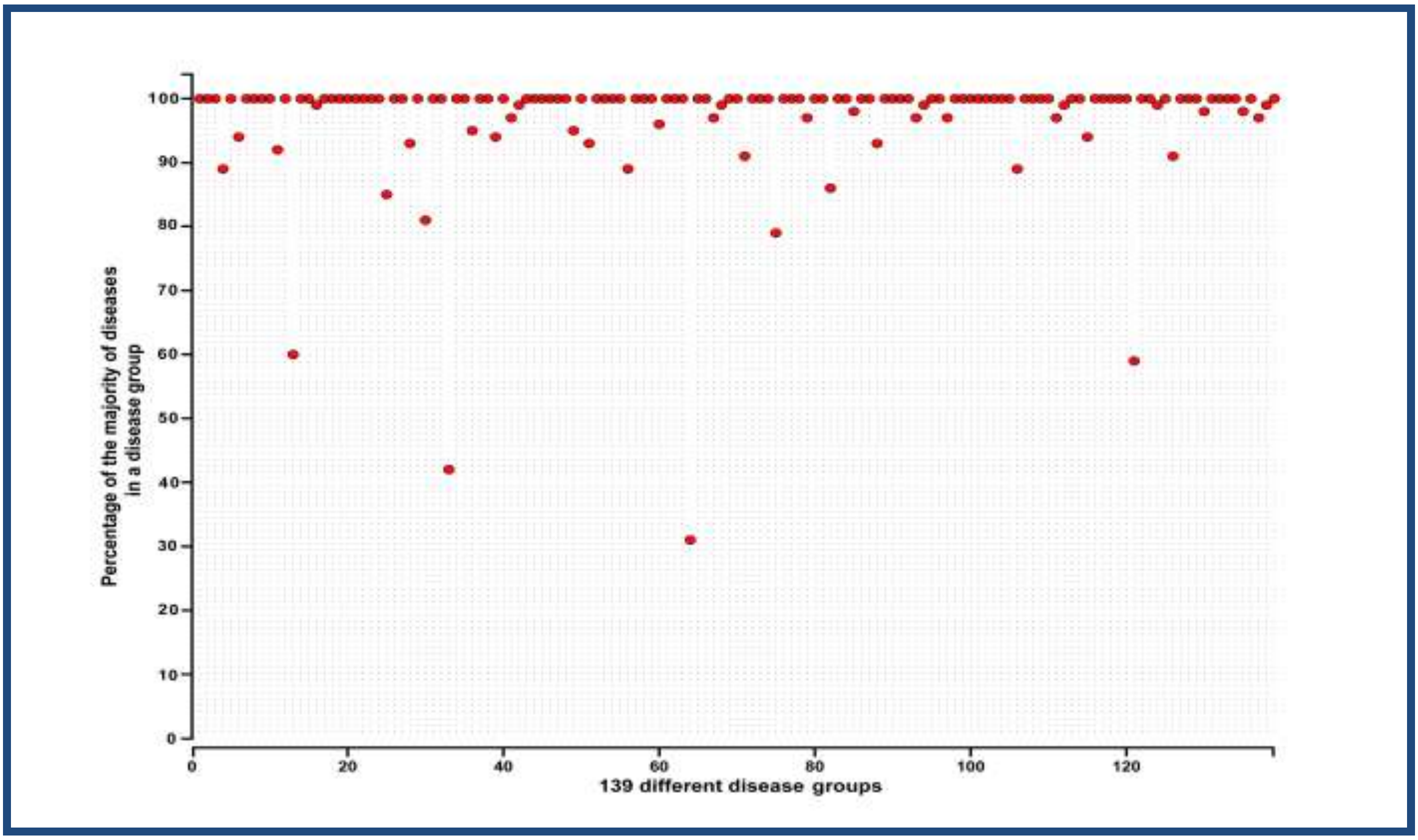

Figure 2: Distribution of the percentage of the majority of disease phenotypes in 139 disease phenotype groups. The diagram shows the distribution of the percentage of the majority of disease phenotypes in 139 different disease phenotype groups. Our result indicated that the corresponding disease phenotypes in the same gene module tend to belong to the same disease class.

Gene modules exaction and evaluation

We used the graph clustering algorithm **[11]** to extract gene modules from HDGN. Here, a module denotes a set of biological individuals (e.g., gene, protein) in a certain biological network, such as gene regulatory network and PPI network. To evaluate the functional relations of genes in gene modules, we introduced the PPI intensity I_{ppi}. I_{ppi} was defined as the fraction of actual existing PPIs among the possible maximum number of PPIs in a gene module *i*, and therefore it can be formulated as:

$$I_{\text{ppi}} = N_{\text{actual}} / (k(k-1)/2)$$

Where N_{actual} is the actual existing PPI number between gene products in gene module, *k* is the number of gene products in this gene module which can be found having interactions with others in the Human Protein Reference Database **[12].**

Disease class enrichment analysis

Disease class enrichment analysis was implemented to explore the enrichment of gene module in a disease class. The framework was executed like this: i) for a gene module, we randomly picked from all the disease genes and built 10,000 pseudo gene modules that have the same number of disease genes as the real gene module , ii) in the real gene module, the possible disease classes of the associated disease phenotypes are determined and the number of disease phenotypes belonging to each certain disease class is counted and iii) the *P*-values for every possible disease class determined in the real gene module are computed based on the random controls.

Results:

HDGN has a highly modular property

We collected all the known 1,865 disease genes from OMIM and constructed HDGN which contains 21,514 links among 1,685 disease genes, with a giant connected component of 1,607 (99.36%) disease genes and 21,428 (98.9%) links **(Figure 1).** In the network, disease gene nodes were marked with different colors based on the assigned disease classes of their associated disease phenotypes. Here, we referred to the disease classification, described by Gol *et al.* **[13],** who manually classified disease phenotypes into 22 main disease classes according to the physiological system affected. It is visually indicative that disease genes with their associated disease phenotypes belonging to the same disease class tend to group together forming different modular structures in the network. **Table 1 (see supplementary material)** listed the two modular measures: dyadicity *D* and heterophilicity *H* for the 22 disease classes. The fact that HDGN has a highly modular structure can be proved by the finding that all the disease classes are dyadic (*D*>1) and most (77.27%) are heterophobic (*H*<1), together indicating a high correlation between the modularity and the physiological classification of disease phenotypes.

Gene products in a gene module tend to interact with each other

Of the total 139 gene modules extracted from HDGN, 14 gene modules have the I_{ppi}=0 because none of their members has interactions with others in the HPRD. Of the remainder, 9 gene modules have the I_{ppi}=1 and the others have the I_{ppi} of 0-1.

Finally, we got the mean I_{ppi} is 0.34. To test the statistical significance of the obtained mean I_{ppi}, 139 gene sets of the same size as the corresponding gene modules were chosen from all the disease genes as a random control. We built 10,000 such random controls and the result showed that the mean I_{ppi} is significantly higher than that of random groups (*P*-value=2.5e-3), indicating that gene products in a gene module have a tendency to interact with each other and be part of the same biological process; that is, these gene products may serve together, as a fundamental functional unit of biological system, to participate in the same cellular pathway or molecular complex.

Gene modules tend to enrich in certain disease classes

We referred to the disease class annotations described by Gol *et al.* **[13]** to conduct disease class enrichment analysis. The result showed that 113 (81.3%) gene modules were significantly enriched in only one of the 22 disease classes, 14 (10.1%) gene modules in two disease classes and 12 (8.6%) gene modules in three or more disease classes. Our statistical results also indicated that the associated disease phenotypes of a given gene module tend to belong to the same disease class **(Figure 2).** Taking together, the vast majority (91.4%) of gene modules have significant specificity to certain disease classes, indicating that these gene modules represent shared genetic origin of the associated diseases, and that genes in a given gene module may be used as a proxy of related diseases in future clinical therapy.

Discussion:

The network modeling method presented here revealed an obvious modular property in HDGN. In the network, the edge between two disease genes represents a measure of their phenotypic relatedness; thus the modularity supports the existing modular organization in genetic diseases **[14],** which is manifested as similar diseases are often caused by functionally related genes. Our findings also showed that disease genes and their associated genetic diseases have a high level of agreement in functional interplay. We believe such functional agreement will prompt the integrative analysis of different levels of biological data. For example, the phenotypic relatedness measure of two genes in HDGN can be considered, combined with gene expression, PPI and GO annotation, to predict candidate genes in an integrated network way.

The measure of gene interactions in HDGN is less quantitative due to the daunting nature of disease phenotype data. In this situation, a weighted HDGN should be considered so that the network is more informative and fit for graph-based clustering algorithm. With increasing amounts of disease phenotype data available, we can construct a more complete map of human disease genes, which make it feasible to investigate the associations among genome, interactome, phenome and other level of omics. We believe these attempts can inform our understanding of the relationship between human diseases and the underlying genetic mechanisms, and further help to uncover pathophysiologic foundations of most genetic diseases.

Conclusion:

We constructed a gene network of 1,865 genes for known diseases called HDGN based on a text-based association determination scheme according to the phenotypic similarity. Disease phenotype data provides a valuable window for dissecting genotype–phenotype associations. Thus, text-based similarity should be a potentially suitable measure for deciding disease gene interactions in HDGN. In addition, HDGN provides a disease-gene-centered sight of disease association map. Hence, it is possible to explore the molecular mechanisms underlying genetic diseases. Genes in 139 gene modules extracted from the network have been demonstrated to functionally interact and the associated disease phenotypes are clinically similar. This observation suggested that related genes cooperate to perform desired cellular functions contributing to certain disease phenotypes. This finds application in target selection and validate during drug discovery.

Acknowledgement:

This research was supported by grant from the Biostatistics Discipline Cultivated Foundation in AHAU (Grant No. 2014XKPY-37).

References:

[1] Freimer N & Sabatti C, *Nat Genet.* 2003 **34**: 15 [PMID: 12721547]

[2] Hamosh A *et al. Nucleic Acids Res.* 2005 **33**: D514 [PMID: 15608251]

[3] Groth P *et al. Nucleic Acids Res.* 2007 **35**: D696 [PMID: 16982638]

[4] Freudenberg J *et al. Bioinformatics.* 2002 **18**: S110 [PMID: 12385992]

[5] Groth P *et al. BMC Bioinformatics.* 2008 **9**: 136 [PMID: 18315868]

[6] Brunner HG & van Driel MA, *Nat Rev Genet.* 2004 **5**: 545 [PMID: 15211356]

[7] Aronson AR, *Proc AMIA Symp.* 2001 **17**: 21 [PMID: 11825149]

[8] Bodenreider O, *Nucleic Acids Res.* 2004 **32**: D267 [PMID: 14681409]

[9] Wilbur WJ & *Yang Y, Comput Biol Med.* 1996 **26**: 209 [PMID: 8725772]

[10] Park J & Barabási AL, *Proc Natl Acad Sci U S A.* 2007 **104**: 17916 [PMID: 17989231]

[11] Bader GD & Hogue CW, *BMC Bioinformatics.* 2003 **4**: 2 [PMID: 12525261]

[12] Peri S *et al. Genome Res.* 2003 **13:** 2363 [PMID: 14525934]

[13] Goh KI *et al. Proc Natl Acad Sci U S A.* 2007 **104**: 8685 [PMID: 17502601]

[14] Oti M & Brunner HG, *Clin Genet.* 2007 **71:** 1 [PMID: 17204041]

Edited by P Kangueane

Citation: **Xuan *et al.*** Bioinformation 11(9): 432-436 (2015)

Supplementary material:

Table 1: Dyadicity *H* and heterophilicity *D* values for the 22 disease categories

Disease class	Disease genes*	In-class links	Out-class links	*D* value	*H* value
Bone	51	112	893	7.4967	0.6145
Cancer	98	769	398	11.145	0.1984
Cardiovascular	101	196	538	5.4931	0.3761
Connective tissue	41	99	1142	12.715	1.4108
Dermatological	102	892	1943	12.174	1.1662
Developmental	48	65	1143	4.1432	1.4103
Ear, nose, throat	54	597	819	49.410	0.6134
Endocrine	79	412	398	7.1568	0.3165
Gastrointestinal	31	121	143	12.140	0.2712
Hematological	69	131	395	5.1562	0.3212
Immunological	59	263	793	10.145	0.4637
Metabolic	203	413	1978	2.3655	0.3431
Multiple	148	1135	2435	6.1359	0.3746
Muscular	62	151	847	4.1283	0.2806
Neurological	249	1897	3986	3.1625	0.5892
Nutritional	27	179	89	38.328	0.2516
Ophthalmological	103	629	798	4.6383	0.4596
Psychiatric	39	237	765	14.568	0.3563
Renal	50	79	254	4.9536	0.5856
Respiratory	35	159	215	29.956	0.3562
Skeletal	56	356	1568	25.562	1.5536
Unclassified	22	14	837	3.9918	1.5077

*The number of disease genes with their corresponding genetic diseases belonging to the left disease category.