TOWARD DISCOVERING DISEASE-SPECIFIC GENE NETWORKS FROM ONLINE LITERATURE

ZHUO ZHANG†, SUISHENG TANG AND SEE-KIONG NG

Institute for Infocomm Research,
21, Heng Mui Keng Terrace, Singapore 119613

Many human diseases are the result of abnormal interactions between multiple genes instead of single gene mutations. Discovering the interactions between these genes and their relationships to human diseases is critical for understanding mechanisms of diseases and helping design effective therapies. Valuable experimental evidence from years of industrious research by biologists can be used to help establish the underlying network of gene interactions related to human diseases. Fortunately, these information are habitually published in research journals whose abstracts are stored in a centralized, easily accessible public database called MEDLINE. To take advantage of this valuable resource, we have developed DiseasePathweaver—a computer-aided knowledge discovery system for extracting, summarizing and visualizing disease-specific gene interaction networks. Using DiseasePathweaver, a biologist can obtain a global overview of the gene interaction network related to a specific human disease, together with well-documented evidences linking to each gene and its putative interactions. We compared the gene networks of two complex human CNS diseases extracted by DiseasePathweaver to the corresponding networks from the human-curated KEGG database and found that our system can accurately cover 79% to 69% of the corresponding disease gene networks, showing the usefulness of DiseasePathweaver as a user-friendly knowledge discovery system for biologists to discover and understand gene interaction networks in complex human diseases. Free access to DiseasePathweaver is available for academic and non-profit users through http://pathweaver.i2r.a-star.edu.sg.

1. Introduction

Many human diseases are caused by multiple genetic and environmental factors. To identify the factors that influence the onset or progression of the disease, a critical step is to unravel the underlying disease-specific gene interaction networks. The current availability of a fully-sequenced human genome, together with the vast biomedical literature that document decades of laboratory and clinical results, and available via a centralized, easily accessible public database called MEDLINE, have opened new avenues for the post-genome scientists to analyze genes implicated in complex human diseases. However, as the majority of clinical research reports and experimentally verified gene interaction information are still habitually stored in unstructured text format in biomedical journals, it is difficult—if not impossible—for a biomedical researcher to unravel the underlying interaction network of the various genes involved in a disease. To do so, bioinformatics tools that can integrate the vast information from diverse resources—including information in unstructured free texts—become highly important for the in-depth study of human diseases.

† contact: zzhang@i2r.a-star.edu.sg

Currently, several disease-related databases are available using various approaches. For example, Gene2Disease (G2D) [1] is a database focusing on pinpointing genes linked to inherited disorders based on assigned chromosomal locations and the possible functional relationships. OMIM [2] is a catalog of human genes and genetic disorders with links to reference and sequence data. MedGene [3] contains information on human gene and disease associations derived from the co-citations of Medline records. Generally, the existing systems have emphasized on direct gene-disease relationship without considering the underlying gene-gene interaction and its network. Indeed, comprehensive gene network construction for understanding complex human diseases is still in its infancy. Only a few molecular interaction network databases are available, for example, KEGG [4]. It contains numerous manually constructed protein interaction networks, mainly in metabolic pathways, and covers only *seven* human diseases to-date. One reason for its small coverage in the interaction network associated with human diseases could be that manually build interaction networks are very difficult to keep up with the speed of current research.

To complement existing disease-related databases in terms of disease-specific gene interaction networks, we have developed a bioinformatics tool called DiseasePathweaver (DPW in short). Our DPW system automates the text mining procedure to extract from relevant documents from Medline, gene interaction and mutation information related to complex human diseases. The system provides a user-friendly web interface for researchers to browse, search, and trace the detailed information in the mined literature. Currently, DPW is focused on human disorders in nervous system and it contains the putative interaction networks for 37 human diseases. We compared here the gene interaction network extracted by DPW to the manually-constructed pathways in KEGG based on two case studies on the Huntington Disease and Amyotrophic Lateral Sclerosis. Our results showed that DPW can cover with accuracy the gene interaction networks for the two complex human CNS (Central Nervous System) diseases that we have studied.

2. Method

2.1. *System Architecture*

As mentioned previously, DPW is an automatic system that is designed to generate disease-specific gene networks based on currently available knowledge on complex human diseases. The system consists of the following components: (1) A disease-related gene extraction module that retrieves relevant information from GeneCards [5]; (2) A text mining module that parses the vast Medline database to extract potentially relevant gene-gene relations and gene mutation data; (3) A relational MySQL database that stores the information that are extracted, including the gene-gene links, gene mutations, their evidences, and the auto-constructed gene networks; and (4) A user-friendly web interface

and visualization module that helps user browse and navigate the networks. Figure 1 shows the detailed system architecture for DPW.

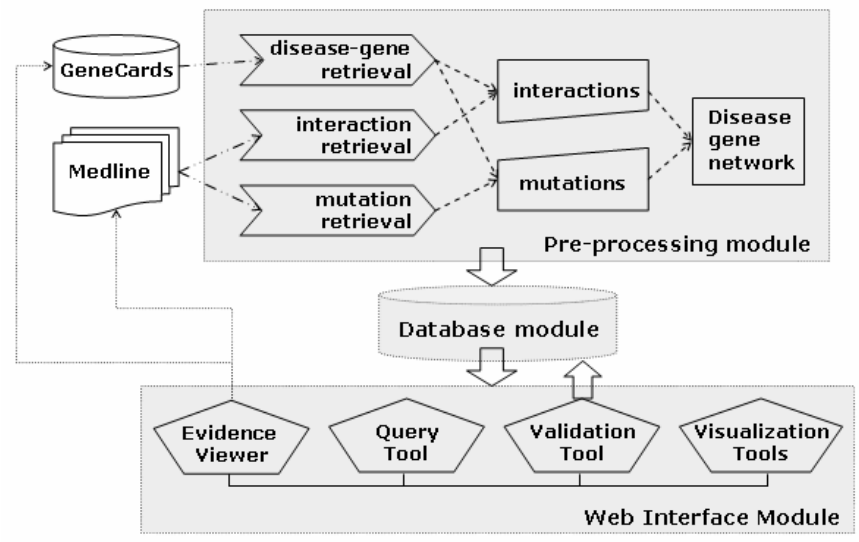

Figure 1. Disease Pathweaver Framework.

2.2. *Procedures*

DPW uses a dictionary of disease names and gene names based on GeneCards [5], an electronic encyclopedia for biological and medical sciences. GeneCards currently contains the information of more than 40,000 human genes, their products and their involvement in diseases. We used this comprehensive dictionary of name identifiers to facilitate literature query in DPW. We call genes included in GeneCards under a disease name as "core genes".

Our main data mining resource is Medline database from NCBI (National Centre for Biological Information). Using a local version of Medline, we were able to exhaustively scan for gene names and related disease names in the literature database. Co-occurrence of genes, as well as co-occurrence of disease and gene mutation information in the abstracts are then collected and deposited into DPW's database. To ensure that appropriate information are extracted and stored, we used an internal scoring system to rank most relevant references extracted.

Next, DPW applies a network building process. Starting from the core genes, DPW searches for possible links and paths between every possible core genes and their working partners, to construct the various disease-specific gene networks. Using only co-occurrence of the "core genes" for a specific disease, the system generates a first

"overview" layer of the related gene networks for that disease. If a "core gene" has interaction partners found from literature that are not specified as a core gene, the system creates a second layer of indirect gene associations. In this way, DPW can present the direct relationships between the respective core genes and diseases in the first layer of an extracted gene interaction network, while further gene interactions related to non-core partners can be displayed in the second layer to provide more insights for the biologists.

In fact, we found that direct links between a pair of core genes do not frequently appear as gene-gene co-occurrence in the abstracts. Possible reasons could be that the direct interaction has not yet been discovered or that the interaction actually occurs through several intermediate steps. In view of this, we enabled DPW to include indirect co-occurrence paths for inferring putative biological interactions between the core genes. Therefore, besides finding co-occurrence genes, DPW also looks for genes that may be involved in intermediate steps as well. The reasoning is, if A interacts with C, and C interacts with B, we can infer a possible biological relationship between A and B through a path A-C-B.

We have implemented our DPW system in a UNIX environment, with data stored in MySQL database. Automated methods for searching the databases and dynamically displaying the selected information and network graphs were built with a combination of Perl, PHP, Java Applet, Graphviz [6] and HTML. In our user-friendly web interface, we provide visualization tools for convenient browsing and retrieving relevant information.

DiseasePathweaver 1.0 currently contains interaction networks for 37 diseases of human Central Nervous System (CNS). We have used 42,620 gene names to scan through whole Medline records (1966~2004). 45,998 pairs of genes co-occur in same abstracts with multiple supporting evidences from the literature were extracted to construct the disease-specific gene interaction networks. Additionally, the system also extracted 491 disease-causing gene mutations with multiple literature evidences. Free access to DiseasePathweaver is available for academic and non-profit users through http://pathweaver.i2r.a-star.edu.sg.

3. Results

In this paper, we compare the automatically extracted disease interaction pathways in DPW with those that were manually curated by biologist experts in KEGG for two CNS diseases, namely the Huntington Disease (HD) and Amyotrophic Lateral Sclerosis (ALS).

3.1. *Case Study 1: Huntington Disease*

Huntington disease (HD) is an inherited, degenerative neurological disease that leads to dementia. The HD gene, whose mutation results in Huntington disease, was mapped to

chromosome 4 in 1983 [7] and cloned in 1993 [8]. With the discovery of the HD gene, new tests were developed that allows those at risk to find out whether or not they will develop the disease. Animal models have also been developed, and we now know that mice have a gene that is similar to the human HD gene. However, research on understanding the mechanisms that cause the HD is still ongoing as it is a complex multifactorial disease.

When we query Medline by typing "Huntington Disease", the search produces more than 5,500 papers to date. It is quite impossible for a biologist to manually sieve through the voluminous amount of literature to discover the interacting network that underlies the disease. Here, we applied DPW's automatic interaction network extraction method as described in the previous section. Figure 2 shows the comparison between the interaction pathways for HD constructed by our automated DPW system and by the manual KEGG experts. The interactions that were found by both KEGG and DPW were depicted with red bold links, the KEGG interactions that were missed by DPW were indicated with green bold lines, while the red dashed lines showed those new putative interactions found by DPW but not by the KEGG expert curators.

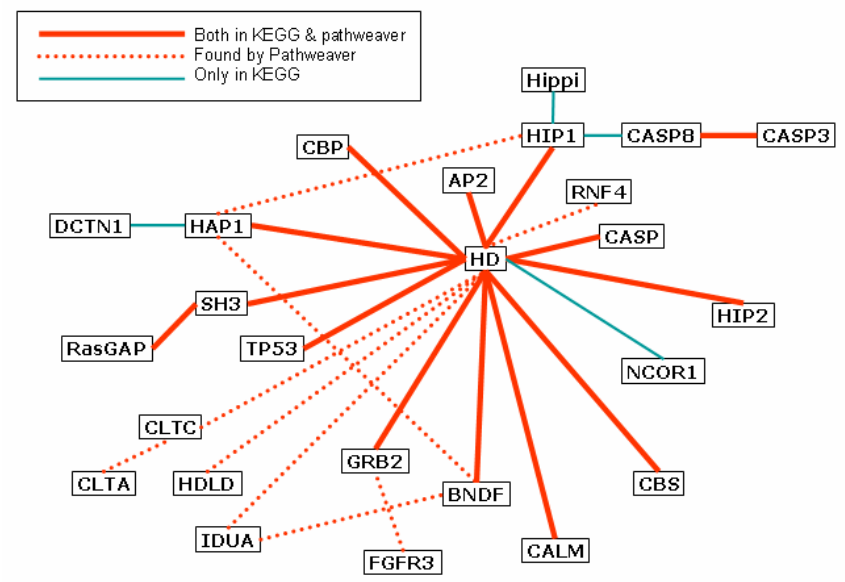

Figure 2. Huntington's Disease Signal Pathway.

The pathway information from KEGG was from http://www.genome.ad.jp/kegg/pathway/hsa/hsa05040.html. Red bold lines indicate interactions found by both KEGG and DPW, green bold lines are the interactions found in KEGG but not extracted by DPW, and red dashed lines show those new interactions found by DPW. (Only part of new interactions are shown here, see online graph for all interactions)

3.2. *Case Study 2: Amyotrophic Lateral Sclerosi*

Amyotrophic lateral sclerosis (ALS) is another neurological disorder characterized by progressive degeneration of motor neuron cells in the spinal cord and brain, which ultimately results in paralysis and death. In 1991, a team of researchers linked familial ALS to chromosome 21 [9]. Two years later, the SOD1 gene was identified as being associated with many cases of familial ALS [10]. The molecular genetics of this relatively new disease is still unclear; a significant amount of research will be required towards promising treatments for ALS.

We have applied DPW to discover a disease-specific interaction network for ALS. The resulting network is shown in Figure 3. Again, we compare the extracted network with the manually curated KEGG pathway for this disease in the figure, which shows that the DPW pathway for ALS is comparable to its manual KEGG counterpart.

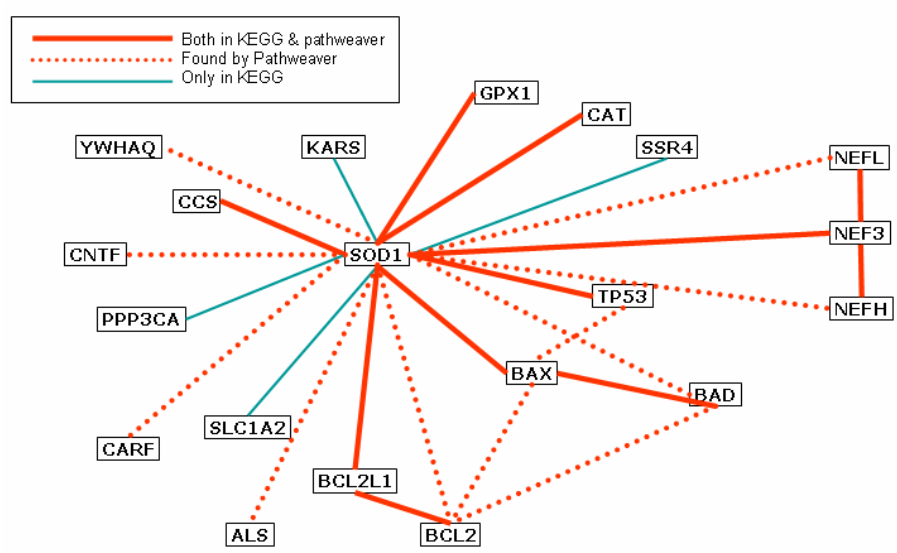

Figure 3. ALS Pathway.

The KEGG pathway information is from http://www.genome.ad.jp/kegg/pathway/hsa/hsa05030.html. The annotation scheme of the interaction graph is same as figure 2.

3.3. *Analysis*

We have analyzed the genes and interactions presented in both pathways. Table 1 shows the numbers of molecules and interactions in both pathways and the coverage percentage of DPW vs. KEGG. In addition to retrieving 79% of the interactions that were also present in KEGG, DPW was also able to generate extra 84% new putative gene

relationships that were not covered by KEGG. Further details of the evaluation can be found in the supplementary data.[†]

Table 1. Coverage of molecules and interactions of DPW

diseases	HD		ALS	
	Molecules	Interactions	Molecules	Interactions
KEGG	18	19	15	13
DPW	38	35	46	54
Coverage	(18/18) 100%	(15/19) 79%	(14/15) 93%	(9/13) 69%

The pathways from KEGG, which was manually drawn by biologists, provided only reference information for each molecule while the information on the interactions is absent. Because of the way the interaction pathways are constructed in DPW, our system is able to provide for each gene, a full list of interactions with its interacting partners, and the interactions are well-documented with the corresponding extracted source of literature. In terms of visualization, a DPW user can easily invoke a graph of the interactions focused on a particular gene of interest. Figure 4 shows how the HD gene is related to other core genes, with some of the interactions are direct, while others links between two genes are through indirect paths involving intermediate interacting partners. The user-friendly interface in DPW allows a biologist to easily validate the putative disease gene interactions by going through the automatically annotated MEDLINE abstracts associated with each extracted interaction in DPW.

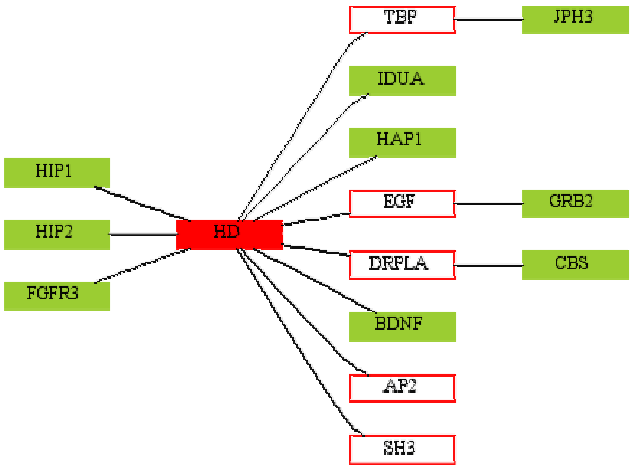

Figure 4. Pathways of HD generated automatically by DPW

[†] Supplementary document is available at http://pathweaver.i2r.a-star.edu.sg/supplementary

The picture was generated by clicking on a gene of interest, which is shown as the red node. Green nodes are other "core genes" related to the disease. White nodes are intermediate genes that link a core gene to another.

In addition to automatically generating disease-specific interaction pathways, the DPW system also extracted gene mutations information and provide relevant links to the source literature. We applied a keyword-driven extraction approach by using a group of keywords such as *mutation, mutated, mutant, deletion, alteration, abnormal, allelic loss, and transversion* to locate candidate sentences that may contain relevant mutation information. We reason that when a sentence contains gene mutation information and also mentions a disease, it indicates a probable link between the mutation and the disease. To avoid irrelevant extraction, we only retained those gene mutation information with multiple literature evidences. The gene mutation information will be crucial for biologists seeking to understand the mechanisms of a disease as well as to design new treatments and cures for it.

4. Conclusion

Most of the current databases on diseases and genes have emphasized on direct relationships of genes and diseases instead of the underlying networks of disease-causing genes and their interaction partners. As we know, genes and their products exercise their biological functions through interaction with other bio-molecules. Analyzing interaction networks is an important step in understanding the mechanisms of physiology and pathology. In this work, we have built an automated bioinformatics system to address the need to cover information about disease-related genes and their underlying interaction networks. Our system, DiseasePathweaver, integrates data extraction, text-mining, automated gene relationship analyses, and user-centric information visualization to facilitate efficient gene interaction research for biologists who are often overloaded with information in the post-genome era. Using Medline as a main reference resource, DPW extracts gene and disease relationship based on the frequent co-occurrence of gene names and disease names in the literature. While this approach does not guarantee true links between genes and diseases, we have shown that they are comparable to manually extracted disease pathways and they also serve well as systematic (and fully automated) frameworks for disease-centric studies. In other words, DiseasePathweaver is an automated system for summarizing and organizing the vast biomedical literature to yield a database of well-annotated disease-specific gene networks that can aid biologists in their study of complex diseases. For further work, we will further improve the quality of DPW's disease interaction networks by constructing a more comprehensive dictionary of biological names and aliases, and applying nature language processing techniques to generate more specific gene interactions more accurately.

References

1. Perez-Iratxeta C, Bork P, Andrade MA. Association of genes to genetically inherited diseases using data mining. *Nat Genet.* 2002 Jul;31(3):316-9. 2002 May 13.
2. Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda,MD),2000. URL: http://www.ncbi.nlm.nih.gov/omim/
3. Hu Y, Hines LM, Weng H, Zuo D, Rivera M, Richardson A, LaBaer J. Analysis of genomic and proteomic data using advanced literature mining. *J Proteome Res.* 2003 Jul-Aug;2(4):405-12.
4. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000 Jan 1;28(1):27-30.
5. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics.* 1998;14(8):656-64.
6. Gansner ER and North SC, An open graph visualization system and its applications to software engineering. *Softw. Pract. Exper.* 1999, 00(S1), 1–5.
7. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY, *et al*. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*. 1983 Nov 17-23;306(5940):234-8.
8. Baxendale S, MacDonald ME, Mott R, Francis F, Lin C, Kirby SF, James M, Zehetner G, Hummerich H, Valdes J, *et al.* A cosmid contig and high resolution restriction map of the 2 megabase region containing the Huntington's disease gene. *Nat Genet*. 1993 Jun;4(2):181-6.
9. Siddique T, Figlewicz DA, Pericak-Vance MA, Haines JL, Rouleau G, Jeffers AJ, Sapp P, Hung WY, Bebout J, McKenna-Yasek D, *et al.,*Linkage of a gene causing familial amyotrophic lateral sclerosis to chromosome 21 and evidence of genetic-locus heterogeneity. *N Engl J Med*. 1991 May 16;324(20):1381-4.
10. Rosen DR, Siddique T, Patterson D, Figlewicz DA, Sapp P, Hentati A, Donaldson D, Goto J, O'Regan JP, Deng HX, *et al*.Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature*. 1993 Mar 4;362(6415):59-62.