Establishment of Disease Pathway Discovery System

Hao-Chang Hsu, Yuan-Chii G. Lee, PhD*

Graduate Institute of Biomedical Informatics, Taipei Medical University, Taiwan
*Corresponding Author
m110094001@tmu.edu.tw, ycgl@tmu.edu.tw

Abstract

We have incorporated genome-wide protein-protein interaction data and transcription regulators data to rebuild a "Disease Pathway Discovery System" (DPDS). This DPDS reconstructs unseen pathways using upstream regulators /downstream regulated genes information from key transcription regulators derived from protein-protein interaction network. We reckoned that if one of the upstream genes goes wrong, many in the cascade of downstream take effect and subsequently amplify the mutant signal by altering the gene expression levels or genetic mutation. These downstream genes are normally easier to be detected as their molecular copy numbers are amplified. Therefore, we aimed to reconstruct the upstream genetic pathway from 2D planar proteins (protein-protein interaction) to 3D transcriptional relationship. We have applied a familial disease Long QT syndrome (LQTS) as case study to test the efficacy of the model. Through this system, we obtained a transcriptional regulatory network from 8 previously reported LQTS related genes. After evaluation with microarray data, this network is confirmed to be heart-exclusive activity. We assume this DSDP is a feasible working model to reconstruct disease pathway.

Key Words: Protein-protein interactions, Protein-DNA interactions, Transcription factor, Gene expression, Pathway discovery

1. Introduction

If one of the upstream genes goes wrong, many in the cascade of downstream take effect and subsequently amplify the mutant signal by altering the gene expression levels or genetic mutation. Many disease-related genes have been announced to be functionally relevant to a certain disease, but as a matter of fact, these genes are most likely playing a part in a downstream of a disease. The reason for that is the gene expression signals of this kind of genes are many fold or even hundred-fold amplified. It is imperative to find a way that can be able to trace the upstream regulatory genes where the disease goes astray. Here in this study, we adopt a genome-wide functional interaction networks, transcriptional regulation networks, and high throughput gene expression profiling of paired normal and disease tissues to construct a platform from where a claimed disease-related gene can be applied to expand its contribution to the disease by extracting its relationship with other protein or its regulatory role in transcription, so that a large-scale, coarse-grained view of cellular networks will be emerged.

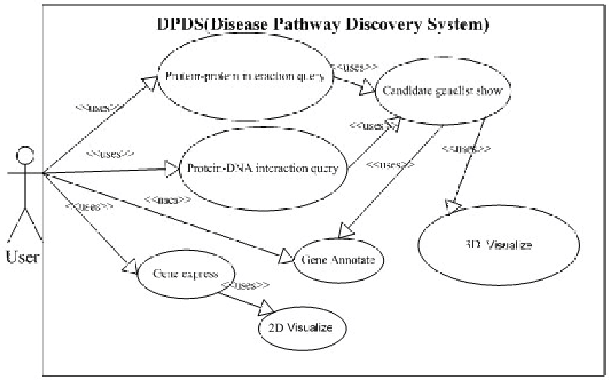

Figure 1 – Use case diagram of DPDS to describe the nature of this system from an external point of view. User may query through gene symbol. DPDS would further allow online visualizing molecular interaction networks.

Protein interactions play a critical role in systems biology of living things. Building up protein-protein interactions not only improves our knowledge towards functional genomics, but also helps predict unknown signaling pathways. High throughput approaches such as yeast two-hybrid system, mass spectrometry, and protein chips are widely applied to extract information of protein interactions in yeast, drosophila, and C. elegance. Nevertheless, the piling up information of protein interactions from PubMed is also invaluable. We will first collect all the available protein interaction databases, such as DIP [1], IntAct [2], MINT [3], BIND [4], HPRD [5], EcoCyc [6] and HIV-1 protein interactions. Among these databases, the last four have been organized, sorted and available in NCBI. Secondly, we will separate the interactions according to different source organisms. Next, we will use Entrez GeneID and Official Gene Symbol (according to the nomenclature of Human Gene Nomenclature Committee, HGNC) as the systematic name for each organism. The reason for this is to facilitate identification of orthologs (among different organisms) and paralogs (in the same organism).

Although protein-protein interactions provide important information towards network connection, yet it lacks time frame to separate the events that might not

occur at the same time. With this regards, we will add time and space scales to the network. Transcription regulation provides such real-time information as one event triggers another in cascade. To reconstitute a complete cellular network it is important to integrate it with transcription regulation data as it is widely accepted that many of the pathways in the cell are regulated both at the transcriptional (protein-DNA) and at the proteomic levels (protein-protein).

To incorporate transcription factors data into the DPDS, we use the data from commercial software, TransPath (www.biobase.com). TransPath® collects information of transcription factors' downstream regulated genes that previously published in the past 30 years. The text formatted data were parsed and then integrated into the DPDS.In addition to the time frame scale, we will add tissue-specific expression information to further depict it actual events to strengthen this network. The microarray data are adopted to fulfill the mission. Su et al., using Affymetrix microarray have made a tissue-specific pattern of transcriptomes to mapping the location for each gene [7]. We took advantage of this gene atlas as space scale.
Finally, the interactive relationship among genes will be visualized using Cytoscape [8], which is a power bioinformatics software for viewing molecular interaction networks.

Through the combination with JWS (Java Web Start) technology, we incorporate successfully Cytoscape into our system, combining back end database with front end virtualization tool. Users can view their network results in the browser directly without further downloading any other tools. The use case diagram for the system can be seen in the Figure 1 and the flowchart of DPDS is in the Figure 2.

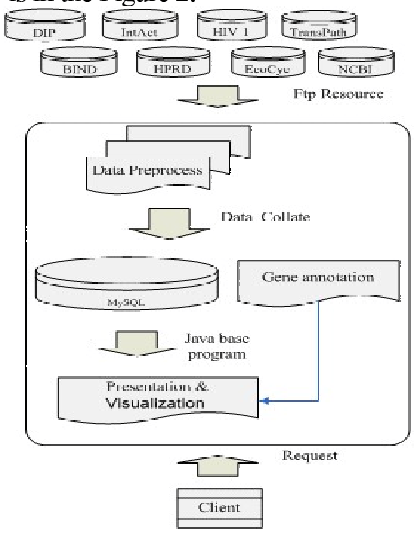

Figure 2 – Flowchart of DPDS (Disease Pathway Discovery System). DPDS is a web base query system to reconstruct disease pathway.

2. Results and Discussion

DPDS is aimed to reconstruct the upstream genetic pathway from 2D planar protein network (protein-protein interaction) to 3D tempo-spatial transcriptional relationship. To prove that our system is feasible and valuable, we take the Long QT syndrome (LQTS) as a case study example. LQTS is a familial disease characterized by an abnormally prolonged QT interval and stress-mediated life-threatening ventricular arrhythmias. As for subtype, there are only eight genes (LQT1 to LQT8) that found to be related to LQTS, which mostly are involved in ion channel regulation [9-14]. We started by the scarce information as only eight related genes, then built a vast protein interacting relationship, and eventually a transcriptional map was emerged. Each gene initially exists individually and finds no connection with each other. Surprisingly, by applying this protein-protein interaction database, the interaction connecting genes are emerged as common modulators and connect one and the other to form a networking web (Level one interactions from 8 LQTS genes are shown in Figure 3 and level 2 interactions in Figure 4). LQTS is though to be a mutation in ion channels, however, through this preliminary data, it provides clue to be associated with neurological synapses, as GRIN1 is shown as one of interacting hub gene. GRIN is a critical subunit of glutamate receptor and a key protein in the synapses, connects LQT1, 2, 5, 6, 7 and 8.

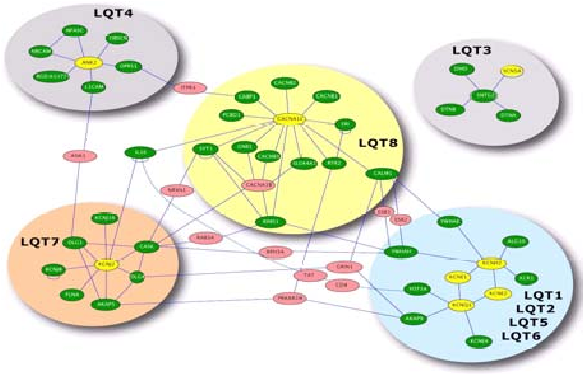

Figure 3 – Protein-protein interaction map facilities understanding the relationship among 8 LQTS genes. The genes in pink background circle Genes in yellow background circle are the original LQTS associated genes, genes in green background circle are the first extended interacting neighbors from the original genes, and genes in pink background circle are the further extended interacting neighbors from genes in green circle.

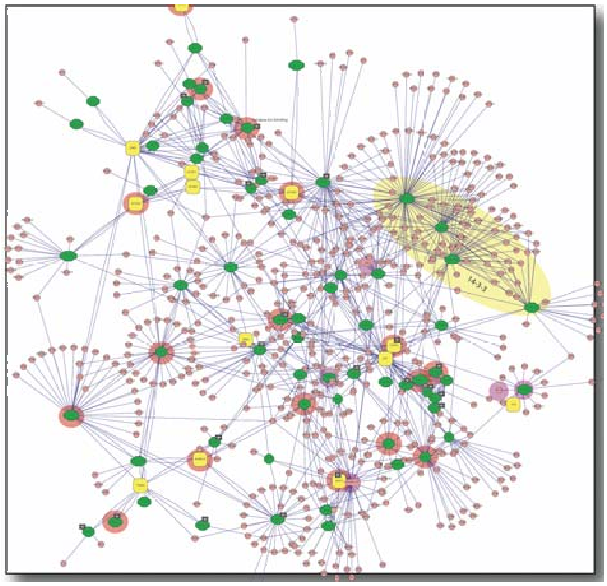

Figure 4 – Further expansion of protein-protein interaction map from 8 LQTS genes. Through DPDS, we can further inquiry associated genes related to Figure 3.

From the protein-protein network map (Figure 4), we selected 7 transcription factors as our studying focus in the sub-project, that is TBX5 (T-box 5), NKX2-5 (NK2 transcription factor related, locus 5/cardiac-specific homeobox), GATA4 (GATA binding protein 4), SMARCA4 (SWI/SNF related, matrix associated, actin dependent regulator of chromatinA4), TCF7L2 (T-cell specific, HMG-box), PAX6 (paired box gene 6), RUNX1 (runt-related transcription factor 1 (acute myeloid leukemia 1; aml1 oncogene). Among these 7 transcription factors, NKX2-5, TBX5 and GATA4 have been reported to be involved in cardiac system, PAX6 is involved in nervous system, SMARCA4 is involved in ATP-dependent chromatin remodeling, RUNX1 is involved in hematopoiesis system. The results are promising as there are 3 out of 7 are already reported to be directly associated with cardiac system. We then extracted the cascade (upstream regulators and downsteam regulated genes) information from TransPath. To demonstrate a clear expression from a complicated network, controlled vocabularies are used to classify the relationship between protein and protein and, protein and gene, such as transactivation, transrepression, phosphorylation activation, phosphorylation repression, activation, repression and DNA binding (Figure 5).

We took advantage of tissue-specific expression microarray data from 79 human tissues [7], as a means to evaluate the lines of evidence in our previously gene predicted network. Among theses tissues, there include heart, atrioventricular node, cardiac myocytes. As shown in Figure 6, since NKX2-5, GATA4 and TBX5 are functioning especially in cardiac system and predicted important transcription factors in cardiac system.

We examined the expression pattern of every single gene including the transcription factors and the regulated genes in the network as shown in Figure 5. The genes are marked with blue background circle, if they are highly or exclusively expressed in heart, atrioventricular node or cardiac myocytes, as shown in Figure 7. There are 40 marked in blue circle out of 89 total gene network, indicating half of the genes presented in this predicted network are highly or exclusively expressed in heart. We conclude that the generated LQTS diseased cellular network using DPDS can be confirmed by the microarray data and will unravel the real unseen cellular world.

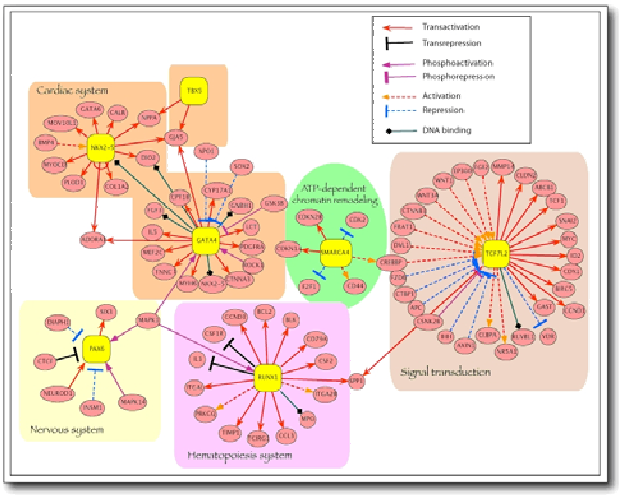

Figure 5 – Regulation of transcription factors in an interlaced web using controlled vocabularies. Seven transcription factors (in yellow square background) were selected from protein-protein interactions associated with cardiac arrhythmias as shown in Figure 4.

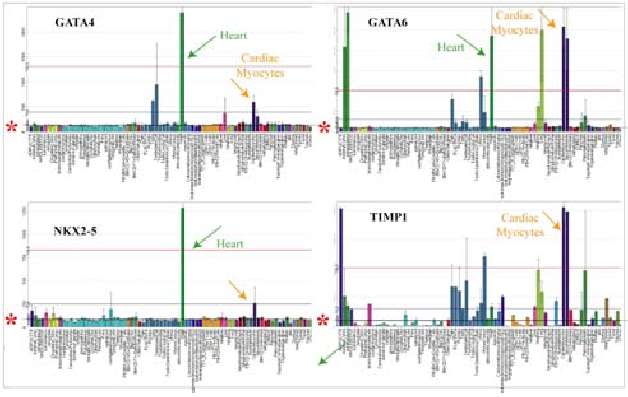

Figure 6 – Gene expression levels in 79 human tissues. The selected genes show highly expressed in heart and/or in cardiac myocytes. The asterisk indicates the median expression value calculated for each gene across all tissues.

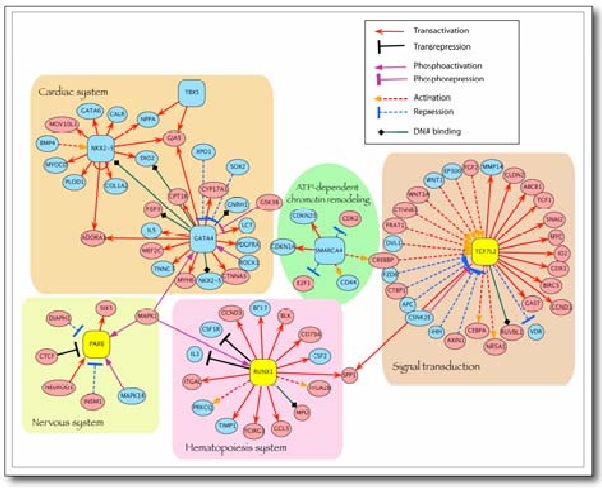

Figure 7 – Genes highly expressed in heart or heart-related tissues are marked in blue background circle in the transcription factor regulatory network. Greater ratio of genes expressed in heart when the transcription factors (shown in the center of each radial network) are highly expressed in heart.

As a means to evaluate the lines of evidence in our previously gene predicted network. Among the different 79 human tissues in Su et al. 2004 microarray dataset there include heart, atrioventricular node, cardiac myocytes. It is reasonable to infer their downstream regulated genes are expressed in higher levels in heart, cardiac myocytes or atrioventricular node. We conclude that the generated LQTS diseased cellular network using DPDS can be confirmed by the microarray data and will unravel the real unseen cellular world.

3. Reference

[1] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Res*, vol. 30, pp. 303-5, 2002.

[2] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler, "IntAct: an open source molecular interaction database," *Nucleic Acids Res*, vol. 32, pp. D452-5, 2004.

[3] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni, "MINT: a Molecular INTeraction database," *FEBS Lett*, vol. 513, pp. 135-40, 2002.

[4] G. D. Bader, D. Betel, and C. W. Hogue, "BIND: the Biomolecular Interaction Network Database," *Nucleic Acids Res*, vol. 31, pp. 248-50, 2003.

[5] S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T. K. Gandhi, K. N. Chandrika, N. Deshpande, S. Suresh, B. P. Rashmi, K. Shanker, N. Padma, V. Niranjan, H. C. Harsha, N. Talreja, B. M. Vrushabendra, M. A. Ramya, A. J. Yatish, M. Joy, H. N. Shivashankar, M. P. Kavitha, M. Menezes, D. R. Choudhury, N. Ghosh, R. Saravana, S. Chandran, S. Mohan, C. K. Jonnalagadda, C. K. Prasad, C. Kumar-Sinha, K. S. Deshpande, and A. Pandey, "Human protein reference database as a discovery resource for proteomics," *Nucleic Acids Res*, vol. 32, pp. D497-501, 2004.

[6] I. M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I. T. Paulsen, M. Peralta-Gil, and P. D. Karp, "EcoCyc: a comprehensive database resource for Escherichia coli," *Nucleic Acids Res*, vol. 33, pp. D334-7, 2005.

[7] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch, "A gene atlas of the mouse and human protein-encoding transcriptomes," *Proc Natl Acad Sci U S A*, vol. 101, pp. 6062-7, 2004.

[8] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Res*, vol. 13, pp. 2498-504, 2003.

[9] M. Csanady and R. Sepp, "[The long QT syndrome from the bedside to molecular genetic laboratory. The history of the first described Hungarian family]," *Orv Hetil*, vol. 146, pp. 2011-6, 2005.

[10] H. Wedekind, T. Bajanowski, P. Friederich, G. Breithardt, T. Wulfing, C. Siebrands, B. Engeland, G. Monnig, W. Haverkamp, B. Brinkmann, and E. Schulze-Bahr, "Sudden infant death syndrome and long QT syndrome: an epidemiological and genetic study," *Int J Legal Med*, vol. 120, pp. 129-37, 2006.

[11] H. Bundgaard, O. Havndrup, M. Christiansen, P. S. Andersen, H. K. Jensen, J. H. Svendsen, and K. P. Kjeldsen, "[Long QT syndrome: genes, mechanisms and risks; indication for genetic family screening?]," *Ugeskr Laeger*, vol. 168, pp. 2537-42, 2006.

[12] D. M. Roden and P. C. Viswanathan, "Genetics of acquired long QT syndrome," *J Clin Invest*, vol. 115, pp. 2025-32, 2005.

[13] L. Crotti, A. L. Lundquist, R. Insolia, M. Pedrazzini, C. Ferrandi, G. M. De Ferrari, A. Vicentini, P. Yang, D. M. Roden, A. L. George, Jr., and P. J. Schwartz, "KCNH2-K897T is a genetic modifier of latent congenital long-QT

syndrome," *Circulation*, vol. 112, pp. 1251-8, 2005.

[14] A. D. Paulussen, R. A. Gilissen, M. Armstrong, P. A. Doevendans, P. Verhasselt, H. J. Smeets, E. Schulze-Bahr, W. Haverkamp, G. Breithardt, N. Cohen, and J. Aerssens, "Genetic variations of KCNQ1, KCNH2, SCN5A, KCNE1, and KCNE2 in drug-induced long QT syndrome patients," *J Mol Med*, vol. 82, pp. 182-8, 2004.