PPINetworkAnalyzer: Revealing the Relationships of Disease Proteins based on Network Analysis Measurements

Sohyun Hwang[1,2] Seung-Woo Son[3] Sang Chul Kim[2] Young Joo Kim[2] Hawoong Jeong[3] Doheon Lee[1]

[1]*Department of BioSystems, KAIST, Deajeon, Republic of Korea*
[2]*National Genome Information Center, KRIBB, Deajeon, Republic of Korea*
[3]*Department of Physics, KAIST, Deajeon, Republic of Korea*
Email : winnie79@biosoft.kaist.ac.kr

ABSTRACT: We made a stepping stone for asthma study by analyzing an asthma-specific protein-protein interaction network. It follows the power-law degree distribution and its hub nodes and skeleton frame of the network agreed with the prior knowledge about asthma pathway. This study is providing a systematic approach to analyze the complex effect of genes or to represent the frame of their relations associated with specific disease.

1 INTRODUCION

Asthma is a major and increasing global health problem and, despite major advances in therapy, many patients' symptoms are not adequately controlled. It is characterized by specific pattern of inflammation in the airway mucosa, and involves the infiltration of eosinophils, increased numbers of T_H2 cells relative to T_H1 cells, and increased numbers of activated mast cells. In addition, there are characteristic structural changes to the airways, some of which might even precede the development of the diseases. These changes include subepithelial fibrosis, airway smooth muscle hypertrophy and hyperplasia, angiogenesis and increased mucus secretory cells. Neural mechanisms are also important in asthma, such as the sensitization of sensory nerve endings in the airways and reflex effects on airway tone. Asthma is a highly complex disease that involves many inflammatory cells, mediators and inflammatory proteins [1].

2 MATERIAL AND METHODS

2.1 Finding candidate genes

First, from Online Mendelian Inheritance in Man (OMIM), the genes associated with asthma were manually found. The OMIM database is known to be more authentic than published papers [2]. Second, from Gene Expression Omnibus (GEO) data of microarray experiments results, the genes associated with asthma were found [3]. All microarray experiments related to asthma were used; Geo Data Set (GDS) 261, GDS266, and GDS267. Experiment of GDS261 (HG_U95a, Affymetrix) is about the comparison of epithelial cells derived from asthmatic and normal airways. Those of GDS266 and GDS267 (HG_U133a and HG_U133b, Affymetrix) are about the investigation of CD4+ lymphocytes from patients with and without atopy, in combination with asthma.

To extract the significant genes associated with asthma from the microarray experiments, Wilcoxon rank sum test was used. Since the microarray data do not follow the normal distribution of assumptions of t test, we adapted Wilcoxon rank sum test which is a nonparametric test for equality of means of two samples that are non-normal. It appears to be a robust choice for microarray data since it operates on rank-transformed data [4, 5]. We took genes of which p-value is less than 0.05.

2.2 Protein-protein interaction network

PPI network associated with asthma was constructed as follows. First, we chose the proteins as nodes of the network corresponding to candidate genes obtained from OMIM and GEO microarray data. Second, we adapted the interactions of candidate proteins as links which were extracted from the Human Protein Reference Database (HPRD) [6, 7]. HPRD provides 22,516 human protein-protein interactions which were found manually by reading the biological research papers [8]. The reliability and quality of the interactions was ascertained by the subcellular localization information of SwissProt.

2.3 Network analysis

To reveal target proteins, which play important roles in the network, several measurements were introduced, for examples the number of nearest neighbors (degree), betweenness centrality (BC), edge BC, and characteristic path length (CPL) of network theory. First, degree distribution gives insight about the existence of hub proteins taking a major role in the network. Second, the BC was measured to find the proteins which are not hub proteins, but take important roles in the view of network topology [9, 10]. BC is a useful measurement for detecting the bottleneck of a network. BC of node k, $b(k)$ is defined as

$$b(k) = \sum_{i,j} b_{i \to j}(k) = \sum_{i,j} \frac{g^k_{i \to j}}{g_{i \to j}}, \qquad (1)$$

where $g_{i \to j}$ is the number of shortest geodesic paths from i to j and $g^k_{i \to j}$ is the number of geodesic paths from i to j that pass through k among $g_{i \to j}$. The edge betweenness centrality is also defined by similar way. Last the CPL is the average number of hopping through shortest geodesic paths from node k to all other nodes. If a node has a small CPL, it means that the node is close to the topological center of the network.

2.4 Functional annotation

To understand the function of target proteins in the view of the biological pathway, Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Database was searched [11]. In addition, to translate lists of candidates into biological phenomenon involved, the information of Gene Ontology (GO) was used [12]. Its statistical significance, p-value, was estimated by adapting the core algorithm of Onto-Express. The p-value can be calculated as

$$p = 1 - \sum_{i=0}^{x-1} \binom{K}{i} \left(\frac{M}{N}\right)^{i} \left(1 - \frac{M}{N}\right)^{K-i}, \qquad (2)$$

where N is the number of genes on the microarray used. N genes are consisted of two categories: functional category (F) and non-F. M is the number of genes included in F among N genes. Let us assume that they picked a subset of K genes. We observe that x of these K genes are in F. N is large enough to change this hypergeometric distribution to the binomial distribution [13].

3 RESULT AND DISCUSSION

3.1 Asthma PPI Network

We constructed two kinds of network. One contains the only asthma associated candidate proteins as nodes corresponding to genes obtained from OMIM and microarray data and their interactions among them as links. The other contains not only candidate proteins but also their interacting proteins as nodes and their all interactions as links. We will refer to the former as core network and the latter as extended network of asthma.

The number of nodes of core network is 337 and the number of edges is 406 and the mean degree is 2.4. There are total 28 clusters with a giant cluster which contains 269 (80%) nodes. The color of each node represents the subcellular location, which ascertains that the interactions are reliable showing the proximity of two interacting proteins (Fig. 1). Among 337 nodes, 79 nodes were found by text manual reading from OMIM database and 264 nodes were extracted by Wilcoxon rank sum test of microarray experiments. 6 nodes were identified by both text manual reading and microarray analysis. The number of nodes in the extended network is 2438 and the number of edges is 4029 and the mean degree is 3.3. The extended network follows a clear power-law distribution, the character of scale-free network (Fig. 2). There are total 78 disconnected clusters with one giant cluster which contains 2256 (92.5%) nodes.

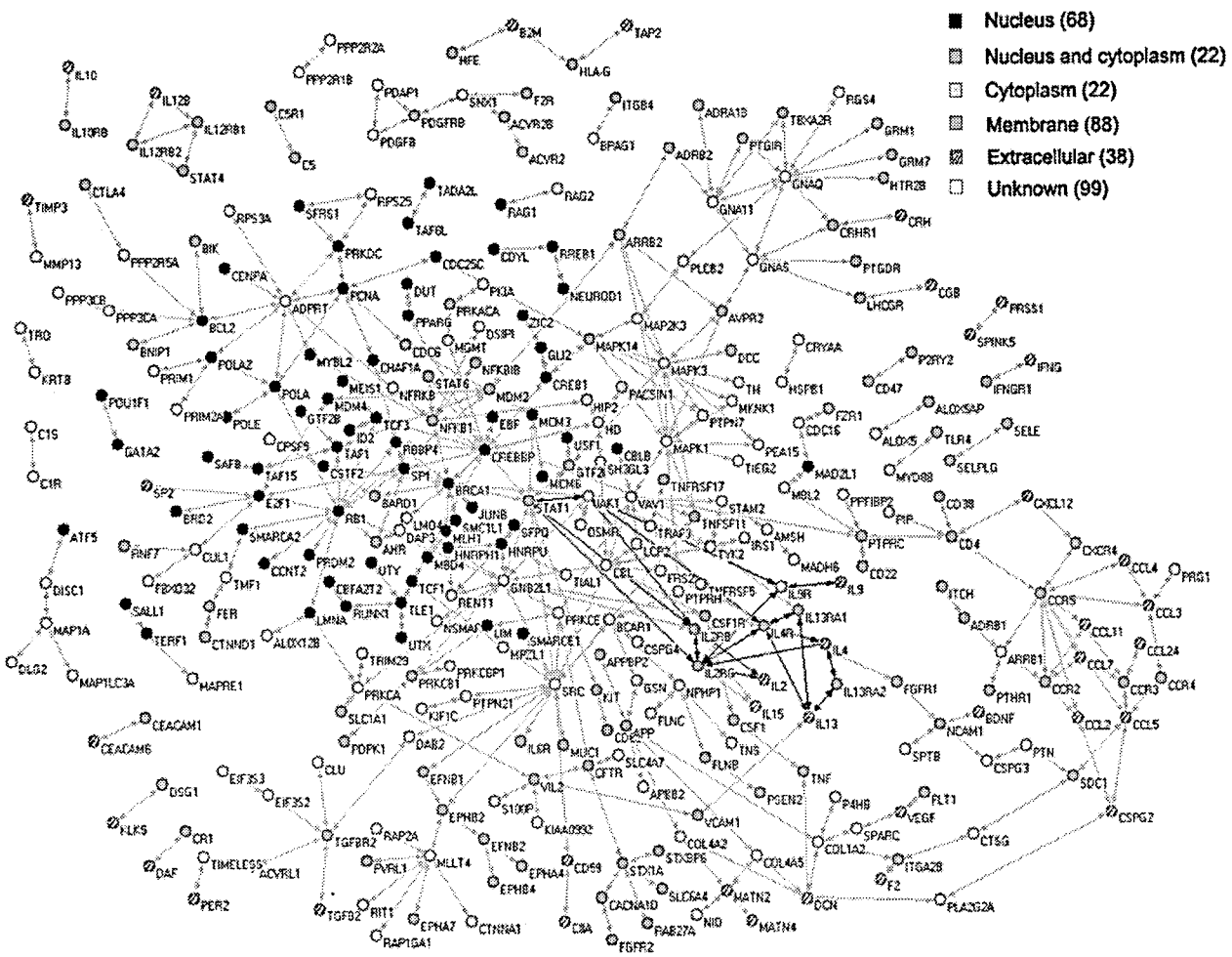

Figure 1: The topological view of the core network of protein-protein interaction of asthma. The subcellular location of its node was represented by its pattern.

Hub nodes appear such as v-src sarcoma viral oncogene homolog (SRC), signal transducers and activators of transcription 1 (STAT1), guanine nucleotide binding protein beta polypeptide 2-like 1 (GNB2L1), retinoblastomal (RB1) and chemokine receptor 5 (CCR5). From the measurement of BC, several nodes have large BC; SRC, GNB2L1, vav1 oncogene (VAV1), Mitogen-activated protein kinase1 (MAPK1), RB1, protein tyrosine phosphatase, receptor type, c (PTPRC), breast cancer anti-estrogen resistance1 (BCAR1) and etc. We confirmed that the hub and large BC nodes are located in the topological center of network by measuring their CPL, verifying central nodes play the key roles in the asthma network.

The hub nodes have many connections to the other part of the network. Therefore they have a large BC and their adjacent links of hub nodes have large edge BC. The *hot* links, which have large edges BC, connect the key nodes of the hubs and large BC nodes as a skeleton [14]. We drew out the sub-network of key nodes from the core network of asthma only using the *hot* links (Fig. 3). The key nodes have an important role to construct the network with hot links. In section 3.3, the biological meaning of them will be discussed.

3.2 Functional annotation of network

First, analyzing the key nodes in the view of biological pathway, they are mostly working for cell communication and cellular process including cell proliferation and cell maintenance (Fig. 3). When searching KEGG pathway with these nodes, MAPK kinase signaling pathway, JAK-STAT pathway, calcium signaling, tight junction, adherens junction, focal adhesion were shown. They are all related to cell communication and signal transduction.

Second, the results of analyzing GO annotation information also agree with previous analysis of key nodes. In all three microarrays, signal transduction (p-value: 4.4E-3, 1.0E-6, 2.6E-4) and cellular physiological process (p-value: 1.9E-4, 1E-5, 1.0E-6) are very significant. In HG_U133b and HG_U95a arrays, metabolism (p-value: 1.0E-6, 1.0E-6) and regulation of physiological process (p-value: 1.0E-6, 7E-5) are significant. In HG_U133b array, neurophysiological process (p-value: 9.7E-4) is significant. In HG_U95a array, development (p-value: 2E-4) is significant.

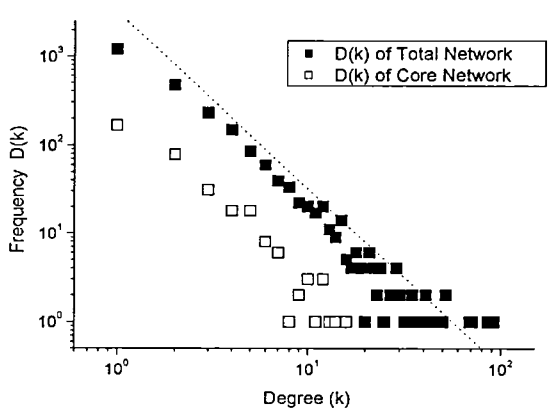

Figure 2: Degree distribution of asthma network

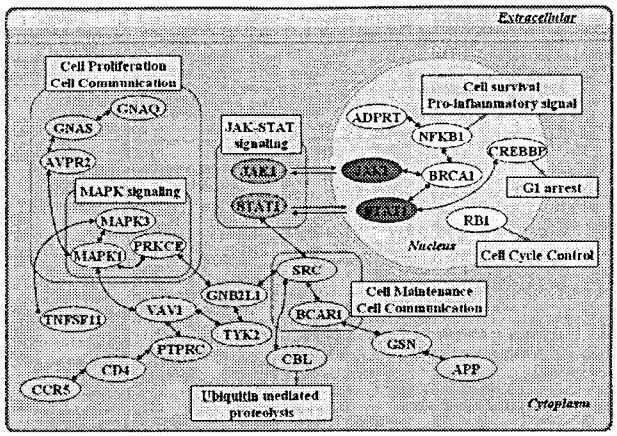

Figure 3: The key nodes of core asthma network. It shows the important skeleton of the core network of protein-protein interaction of asthma. One node represents a protein, but the node name follows the gene's name for its convenience.

3.3 The roles of key nodes in asthma-pathway

The JAK-STAT signaling pathway provides one of the most direct routes from cell-surface receptors to the nucleus. It is intimately related with the effects of interferons, hormone and interleukins. Through JAK-STAT pathway, several genes related to inducing inflammation can be recruited. In the extended network of asthma, interleukin (IL) 4, IL13, IL9 and IL2 through their receptors interact with Janus Kinase (JAK) 1, which is a component of JAK-STAT pathway (Fig. 1). IL4, as the proinflammatory cytokine, induces the eosinophilic inflammation and is to promote differentiation of T_H2 cells, acting at a proximal and crucial point in the allergic response. IL4 and the closely related cytokine IL13 signal through a shared surface receptor, IL4Rα. IL13 has several effects relevant to allergic inflammation in asthma including production of immunoglobulin E (IgE) from B lymphocytes. IL9 is the T_H2 cytokine that enhances T_H2-driven inflammation, amplifies mast-cell mediator release and IgE production, and enhances mucus hypersecretion [1].

In nucleus of Fig. 2, three genes control the cellular proliferation. Nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (NFKB1) is pro-inflammatory signaling molecules and is related with survival. cAMP responsive element binding protein binding protein (CREBBP) has a role controlling the G1 arrest in the cell cycle and RB1 also controls cell cycle. In cytosol, SRC and BCAR1 work for cell maintenance in the focal adhesion pathway. Ubiquitin mediated proteolysis is working through Cas-Br-M ectropic retroviral transforming sequence (CBL). MAPK signaling pathway and SRC are also working for cell communication [15].

ACKNOWLEDGEMENTS

This work was supported by National Research Laboratory Grant (2005-01450) from the Ministry of Science and Technology. SWS and JH are supported under grant No. R14-2002-059-01002-0 from the KOSEF-ABRL program and SH, KSC and KYJ are supported under the grant from

the Stroke Oriental Medicine Project (M1052701000005N270100000) of the Ministry of Science and Technology of Korea.

REFERENCES

[1] P. J. Barnes. New drugs for asthma. *Nature Rev. Drug Discovery*, 3, 831-844, 2004.

[2] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, V. A. Mckusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33, D514-D517, 2005.

[3] R. Edgar, M. Domrachev and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30, 207-210, 2002.

[4] O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein, R. B. Altman. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18, 1454-1461, 2002.

[5] R. E. Walpole and R. H. Myers. Probability and statistics for engineers and scientists., 5th ed, Macmillan, New York, 1993.

[6] D. Eisenberg, E. M. Marcotte, I. Xenarios, I. O. Yeates. Protein function in the post-genomic era. *Nature*, 405, 823-826, 2000.

[7] H. Jeong, S. P. Mason, A. L. Barabasi, Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411, 41-42, 2001.

[8] 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, A. Pandey. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Research*, 32, D497 - D501, 2004.

[9] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40, 35-41, 1977.

[10] M. E. J. Newman. Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E*, 64, 016131, 2001.

[11] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27, 29-34, 1999.

[12] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, R. White. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32, D258-D261, 2004.

[13] S. Drăghici, P. Khatri, R. P. Martins, G. C. Ostermeier, S. A. Krawetz. Global functional profiling of gene expression. *Genomics*, 81, 98-104, 2003.

[14] D. H. Kim, J. D. Noh and H. Jeong. Scale-free trees: The skeletons of complex networks. *Phys. Rev. E*, 70, 046126, 2004.

[15] Alberts, B. The molecular biology of the cell. Garland Science, 2002.