Mining a Large-scale Microarray Database for Similar Gene Expression Modules to Find Distant Relationships between Down Syndrome and Huntington's Disease

Yoshifumi Okada and Wataru Fujibuchi*
Computational Biology Research Center, Advanced Industrial Science and Technology (AIST), Japan
*Corresponding author: w.fujibuchi@aist.go.jp

Abstract

Sequence motif search is one of the main themes of data mining approaches in bioinformatics. We have developed a new technology for searching a large-scale of gene expression data (typically contains thousands of experiments) for similar gene expression patterns or motifs (called `gene modules') to a query experiment. The program is reasonably fast and usually outputs the exhaustive search results within a minute. The resulted gene modules consist of a subset of genes and a subset of experiments. Using the disease and syndrome database, we made an exhaustive dictionary of modules by querying with each disease experiment. Through the module analysis we find novel relationships between distant diseases that are not previously identified. For example, modules found in Down syndrome and Huntington's disease indicate involvement of cell adhesion molecules in both cases, which could suggest common mechanisms of causing the similar phenotypes, neural and muscle disorders, observed in the clinical data.

Background

The number of microarray experiments available in GEO (1) database is growing almost exponentially every year. Although such a huge resource of gene expression data is available, few of data mining tools have been developed to extract useful biological information so far. The BLAST-like algorithm in the microarray field has been expected to appear since the early years of microarray analysis (2) but it has not been realized due to its extremely large calculation space.

In recent years, a new clustering method called biclustering where common gene expression patterns are clustered with regard to exhaustive combinations of experiments has been devised and widely studied (3,4, etc.). Okada *et al* have developed the fastest algorithm for exhaustive search of biclusters (5) and successfully enumerated the maximal biclusters, or gene modules that demonstrate the highest enrichment of gene function sets among five tested software programs in yeast data.

In this paper, we extend their algorithm to a database search tool as well as further enhance the search quality by improving the filtering process of enumerated modules. Using the disease and syndrome microarray database, we demonstrate the usefulness of the method by showing that the extracted gene modules correspond well to known biological networks.

Mining a database for modules related to a query experiment

We have developed a heuristic algorithm for extracting exhaustive gene modules that share common gene expression patterns in both of query and database. **Figure 1** is a schematic view of the entire process of the search.

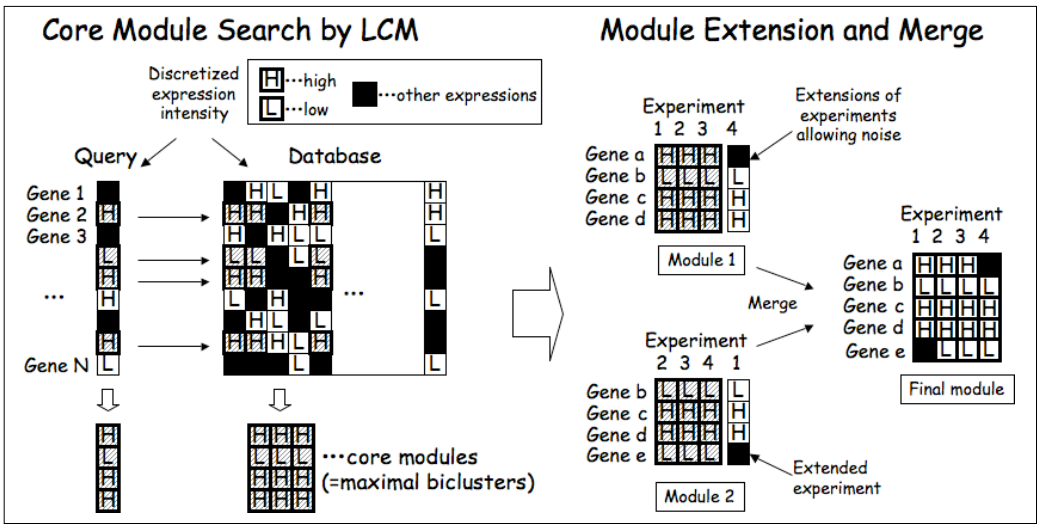

Figure 1. *Schematic view of extracting gene modules from database. Left: the gene expression data are discretized and searched for core*

gene modules by LCM algorithm. Right: modules are merged by incorporating more experiments containing noised data.

The details of its algorithm will be published somewhere else thus we will describe a brief outline of the scheme in this paper. Firstly, the gene expression values in database and query are transformed to rank orders within each experiment. Secondly, rank orders of each gene are further discretized to either of `high', `low', and `others' according to the rank order distribution in the database. In this study we assign `highs' for the top 0.5% of rank orders and `lows' for the bottom 0.5% for each gene. Thirdly, we apply the fastest data mining algorithm called 'Linear time Closed itemset Miner (LCM)' (6) to extract core gene modules that display exactly the same expression patterns between the query and database. Lastly, the core modules are extended to experiment direction allowing mismatches of values and merged each other if possible. This merge process effectively reduces the number of similar but redundant modules by 20~50% of the initially extracted core modules.

Preparation of normal, cell-line, and disease microarray database

First, we manually organize normal, cell-line, and disease microarray databases from the provided dataset. Then, for each database we performed all-to-all microarray data similarity search by our fast Spearman rank correlation search method (7) to eliminate data redundancy and create non-redundant databases. We reduce 1,290, 1,139, and 3,467 data to 724, 345, and 2,899 for normal, cell-line, and disease databases, respectively. For each database, we perform module searches by querying each data using the above method. Hereafter, we will focus only on the disease database and describe the analytical results in detail.

Analysis of disease-related database

The overall statistics of extracted modules searched in disease database by querying each disease data are shown in **Figure 2**. The final numbers of gene expression modules obtained in this method vary depending on the query, ranging zero to thousands. Thus, we used only the top 100 modules containing the largest experiments for each query for evaluation. To analyze how significantly genes in modules are involved in particular biological functions, we compare modules with gene ontology (GO) terms and KEGG pathway maps, and calculate the significance of function enrichment for each module using hyper-geometric distribution probability.

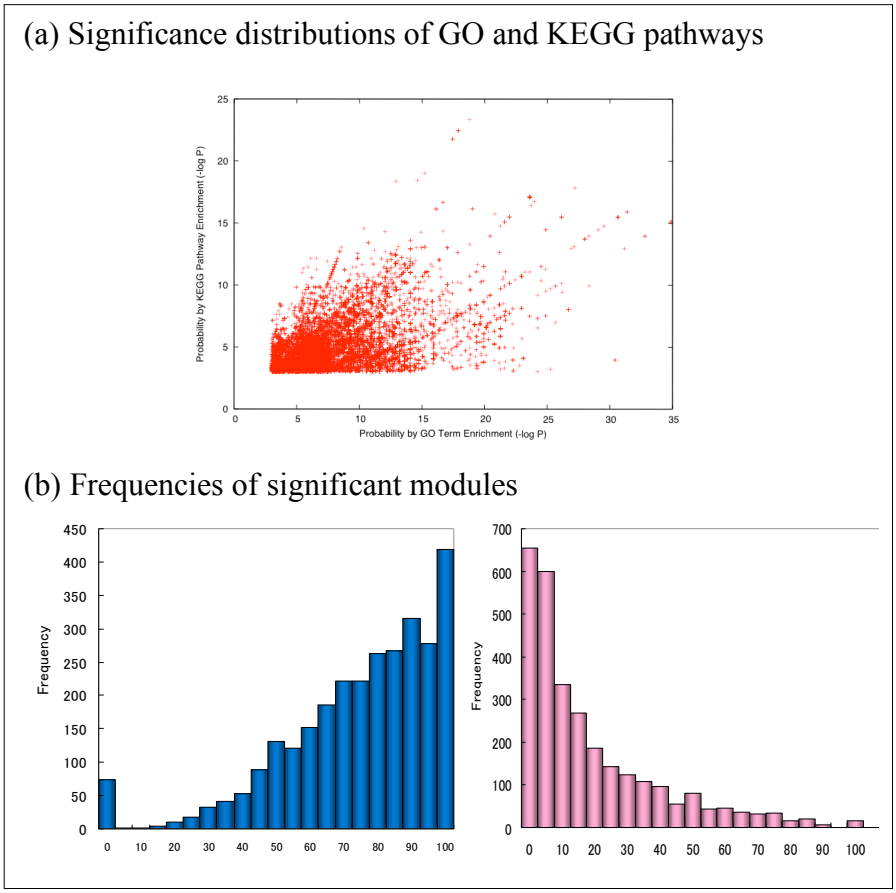

Figure 2. *Statistics of gene modules in disease database. The top 100 of the largest modules for each query are statistically examined and shown: (a) their modules with probabilities of gene function enrichment in GO (horizontal axis) and KEGG pathway maps (vertical axis); and (b) frequencies of queries against the ratio of their significant modules ($p<0.001$) in GO (left) and KEGG pathway maps (right).*

Statistically, the number of significant modules is much larger than that of modules expected with the random genes and is correlated between GO terms and KEGG pathways. In

this analysis, many of extracted modules tend to represent functionally enrichment in more GO terms than KEGG pathway maps. For example, all modules extracted with more than 400 of queries are completely matched with GO terms under the threshold ($p<0.001$), while those with less than 20 queries are completely matched with KEGG pathway maps. This is probably due to the differences of gene members in single GO terms and KEGG pathway maps; there are only a few to tens of members in single GO terms, while tens to hundreds in single KEGG pathway maps, thus it is hard to obtain high significance from KEGG pathway data.

Gene modules from five diseases and their relationships

To scrutinize the extracted modules, we pick up five diseases, Down syndrome (DS), Huntington's disease (HD), T-ALL (TL), B-ALL (BL), and Myotrophic lateral sclerosis (MLS) from the database and show their module numbers before and after merging process, GO and KEGG functional enrichment significances, and their few examples of highly significant modules in **Table 1**.

Disease Name	#Core Modules	#After Merging	Ratio of GO Significant Modules in top 100 ($p<0.001$)	Ratio of KEGG Significant Modules in top 100 ($p<0.001$)	Examples of Highly Significant Modules ($p<1e-10$)
Down syndrome	211	48	86%	38%	Cell adhesion, Intermediate filament cytoskeleton organization and biogenesis, Wnt receptor signaling pathway through beta-catenin, beta-catenin binding, striated muscle contraction
Huntington's disease	2031	868	89%	4%	Ammonium transporter activity, prostaglandin biosynthetic process
T-ALL	413	235	48%	4%	–
B-ALL	449	206	60%	4%	Prostaglandin E receptor activitiy
Myotrophic lateral sclerosis	5385	1452	89%	40%	Cytosolic small ribosomal subunit, structural constituent of ribosome, integrin binding, sulfonylurea receptor activity, Intracellular, GTPase activator activity

Table 1. *Five examples of diseases used for queries and their extracted modules. Only the top 100 modules containing the largest experiments are tested their gene function enrichment for avoiding spurious modules that are little preserved across the experiments.*

As shown in these examples, core modules are reduced to 20-50% in size by merging process to generate final modules. The ratios of significant modules in GO terms and KEGG pathway maps are 48–89% and 4-40%, respectively. Note that as high as 86-89% of tested modules of DS, HD, and MLS are significantly matched with GO terms. This percentage is extremely significant because this ratio dramatically decreases when gene ids are randomly shuffled within microarray data; for example, 86% of DS decreases to 72% for $p<0.001$ and 71% decreases to 23% for $p<0.0001$ (average in 10 random shuffles).

Overlapping GO terms and KEGG pathways in different diseases

To detect unknown relationships among diseases we have investigated in overlapping GO terms and KEGG pathways between all pairs of five diseases. Under a substantially small threshold of probability ($p<1e-6$ for both diseases) we find three GO term overlaps for DS vs. HD, two overlaps for HD vs. BL, TL vs. MLS, and BL vs. MLS. We also find one KEGG pathway overlaps for DS vs. BL, and TL vs. MLS.

Cell adhesion molecules involved in both Down syndrome and Huntington's disease

Among overlapping GO terms between different diseases, all of the three GO terms found in both DS and HD are related to cell adhesion molecules and the integrin-mediated signaling pathway. According to published documents, the cell adhesion molecules called DSCAM (Down syndrome cell adhesion molecule) and NCAM (neural cell adhesion molecule) are strongly involved in the neural developments and disorders both in DS and HD, respectively (8,9).

Discussion and Conclusion

Both DSCAM and NCAM are members of the immunoglobulin superfamily cell adhesion molecules (IgCAMs) (10). Quite interestingly, DSCAM is also proposed to cause congenital heart disease in DS patients (8), while patients of HD develop severe hyperkinetic motor disturbances triggered by yet unknown molecular events (9). Taken together, we speculate that NCAM be a possible candidate to cause part of motor disorders in HD.

In this paper, we have shown our novel gene module search method that can enumerate all possible core modules with regard to query experiments and tune them by allowing noise. The resulted modules are significantly involved in particular GO terms and KEGG pathway maps. Among modules, we find three modules that share the common GO terms in two unrelated diseases, DS and HD. Investigating the underlying molecular mechanisms, we find that IgCAMs are highly involved in both diseases in terms of neural disorders. Thus we conclude that the microarray data mining based on our gene module search with queries is highly useful for extracting new links and distant relationships that are currently unrecognized among different diseases.

References

1. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, *et al.* (2007), Nucleic Acids Res. 35:D760-5.
2. Basset, Jr., D., Eisen, M.B., and Boguski, M.S. (1999), Nature Genet. 21:51-55.
3. Cheng, Y. and Church, G. (2000), ISMB2000:93-103.
4. Tanay, A., Sharan, R., Kupiec, M., and Shamir, R.(2004), Proc. Natl. Acad. Sci. U.S.A. 101:2981-2986.
5. Okada, Y., Horton, P., and Fujibuchi, W. (2007), IAENG International Journal of Computer Science 34:119-126.
6. Uno, T., Kiyomi, M., and Arimura, H. (2004), IEEE ICDM'04.
7. Fujibuchi, W., Kiseleva, L., Taniguchi, T., Harada, H., and Horton, P. (2007), Bioinformatics (accepted).
8. Barlow, G.M. *et al.* (2001), Genet Med. 3(2):91-101.
9. van der Borght K, and Brundin P. (2007), Exp Neurol. 204(1):473-8.
10. Fusaoka E, Inoue T, Mineta K, Agata K, and Takeuchi K. (2006), Genes Cells 11(5):541-55.