Report from microarray data analysis contest (CAMDA 2007)

Yoshifumi Okada

Assistant Professor, Computer Science and Systems Engineering, Muroran Institute of Technology
former AIST Research Staff, Cell Function Design Team (until March 31st, 2008)

The Critical Assessment of Microarray Data Analysis (CAMDA)[1] competition was established in 2000, and convenes annual microarray data analysis contests in which contestants compete in a two-step process involving 1) analysis of the challenge datasets and the submission of papers and 2) evaluation by CAMDA judges. Teams judged to have shown particular merit are invited to give oral presentations. Unlike the CASP contest involving protein structure predictions, the results in the CAMDA competition have not been experimentally validated, and evaluations are based on both analytic method and biological findings.

In December 2007, I participated in the CAMDA 2007 event in Spain along with Team Leader Fujibuchi of the Cell Function Design Team. Our team opted to employ our gene module search system SAMURAI[2] to analyze the large-scale challenge data set consisting of expression data for various diseases and syndromes. We constructed a 22,283 gene × 2,899 array database using these data, with the goal of searching for gene modules representing minimal units of genes specifically co-expressed in disease cells and uncovering various biological relationships across diseases. Module searches, however, involve the searching of combinations of genes and arrays (diseases), meaning that extant search methods would be limited to a few thousand genes over a few hundred arrays due to its high computation complexity. In order to optimize SAMURAI for such large datasets, we extended it to more high-speed and powerful program that can dramatically reduce the search space at the point of query by excluding genes that could not be a component of a gene expression module[3] (Fig.1). By querying our database using individual array data, we found that for nearly all queries we were able to systematically discover gene modules in only a few seconds to a few minutes. Functional analysis of our results further revealed statistically significant functional relationships between Down syndrome and Huntington's disease in both cell adhesion and cell surface recognition characteristics. Interestingly, molecules known to be involved in these functions have been shown in *Drosophila* and mouse to play important roles in the formation and impairment of neural circuits, and in humans may be associated with the mental disability seen in Down syndrome and Huntington's disease.

These findings were rated highly by the CAMDA judges, and were selected for an oral presentation. I feel that our participation in the CAMDA competition was a wonderful opportunity to demonstrate the power of SAMURAI technology to the global research community. In the future, we seek to add a function for detecting temporal changes in expression patterns of modules, an evolutionary advance that would allow, for example, for the discovery of dynamic processes of development, differentiation or pathogenesis, in preparation for the next CAMDA meeting (Austria).

References

1. http://camda.bioinfo.cipf.es/
2. Okada, Y., Fujibuchi, W. and Horton, P., "A biclustering method for gene expression module discovery using closed itemset enumeration algorithm", IPSJ Transactions on Bioinformatics, vol. 48, no. SIG 5(TBIO2), pp.39-48, (2007).
3. Okada, Y. and Fujibuchi, W., "Mining a large-scale microarray database for similar gene expression modules to find distant relations between Down syndrome and Huntington disease", The 7th Int. Conf. for the Critical Assessment of Microarray Data Analysis (CAMDA 2007), Valencia, Spain, (2007).

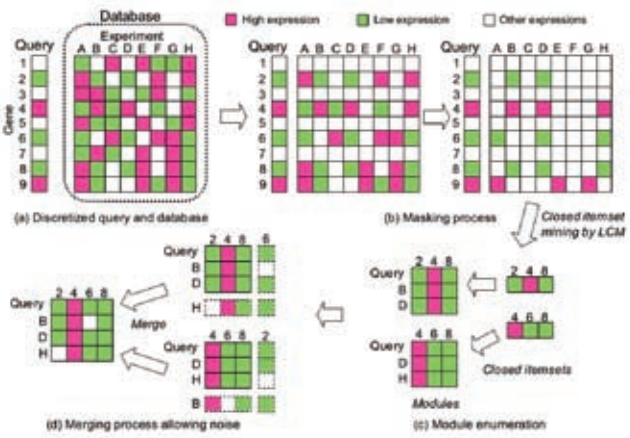

Fig.1 : SAMURAI processing method

(a) Advance discretization of query and database values
(b) Genes which cannot form modules are excluded so as to allow for the searching only of modules in which queries show high or low expression
(c) Closed itemset enumeration called LCM algorithm is used to enumerate sets of genes forming modules
(d) Similar modules are merged allowing noise