Bioinformatics Center
- Pathway Engineering -

http://www.bic.kyoto-u.ac.jp/pathway/index.html

Prof
MAMITSUKA, Hiroshi
(D Sc)

Assist Prof
TAKIGAWA, Ichigaku
(D Eng)

Assist Prof*
SHIGA, Motoki
(D Eng)

Program-Specific Res
KAYANO, Mitsunori
(D Sc)

Program-Specific Res
NATSUME, Yayoi
(D Agr)

Program-Specific Res
HANCOCK, Timothy Peter
(Ph D)

PD (JSPS)
NGUYEN, Hao Canh
(Ph D)

*Bioinformatics Center, Laboratory of Advanced Data Mining for Bioinformatics.

Student

du VERLE, David (D1)

Visitors

Dr ZHU, Shanfeng	Fudan University, China, 22 January–6 February 2009
du VERLE, David	Inserm, France, 1 February–17 March 2009
NGUYEN, Hao Canh	JAIST, Japan, 23 February 2009
Dr NG, See-Kiong	A*Star, Singapore, 16–23 April 2009

Scope of Research

With the recent advancement of experimental techniques in molecular biology, research in modern life science is shifting to the comprehensive understanding of a biological mechanism consisting of a variety of molecules. Our focus is placed on molecular mechanisms in biological phenomena, represented by biological networks such as metabolic and signal transduction pathways. Our research objective is to develop techniques based on computer science and/or statistics to systematically understand biological entities at the cellular and organism level.

Research Activities (Year 2009)

Publication

Kayano M, Takigawa I, Shiga M, Tsuda K, Mamitsuka H: Efficiently Finding Genome-wide Three-way Gene Interactions from Transcript- and Genotype-Data, *Bioinformatics*, **25 (21)**, 2735-2743 (2009).

Presentations

Mining Significant Patterns from Trees, Mamitsuka H, Université Louis Pasteur, Strasbourg, France, 28 May 2009.

Clustering with Heterogeneous Data, Mamitsuka H, IEEE International Conference on Computational Intelligence and Natural Computing (CINC 2009), Wuhan, China, 6 June 2009.

A Markov Classification Model for Metabolic Pathways, Mamitsuka H, Fudan University, Shanghai, China, 29 September 2009.

Mining Significant Patterns from Glycan Structures, Mamitsuka H, International Beilstein Symposium on Glyco-Bioinformatics, Potsdam, Germany, 5 October 2009.

Grants

Mamitsuka H, Integrative Data Mining Approaches for Unstructured Data in Life Sciences, Research Grant from BIRD (BioInformatics Research and Development) of JST (Japan Science and Technology Agency), 15 October

Efficiently Finding Genome-wide Three-way Gene Interactions from Transcript- and Genotype-Data

The topical work this year is the issue of finding a three-way gene interaction, precisely two interacting genes in expression under the genotypes of a different gene, given a dataset in which both gene expressions and genotypes are measured for each individual. We illustrate our problem setting by using synthetic 2D diagrams in Figure 1, where expression values of two genes are plotted with three classes (genotypes): +, * and Δ. In this figure, (a) shows expression values being just randomly distributed; (b) shows expression values being easily categorized into three classes; and (c) shows that classes can be categorized by expressions without using two genes at the same time. We are not interested in (a–c) but in (d), which shows that the correlation in expression between two genes differs for each class. More concretely, two genes are positively correlated for one class, whereas they are negatively correlated for another. This is exactly a switching mechanism in expression between correlation and inverse-correlation of two genes, controlled by another gene. At the same time, this is the three-way gene interaction which we are interested in. We emphasize that this interaction is key to elucidating complex biological systems. A usual, common approach to detect the three-way interaction is the likelihood ratio test for regression. Particularly, logistic regression must be suitable the most, because of categorical responses (genotypes) in our setting. However, parameter estimation for logistic regression is based on the maximum likelihood, for which only a time-consuming iterative gradient descent, Newton–Raphson, is usually used. In our case, classes are genotypes, resulting in a problem of an explosive number of combinations of one SNP (genotypes) and two genes (expressions). For example, for 50,000 SNPs and 1,000 genes, we have roughly 5×10^{10} (= 50,000×1,000×1,000) combinations, making scanning over all possible combinations intractable. Thus, the main focus of this work is to speed up the procedure of finding the three-way interactions. Our strategy for this issue is to prune irrelevant combinations, such as those in which the expression values of two genes are randomly distributed as in Figure 1(a), by using statistical testing assuming the normality of given examples. Our experiments with a huge dataset of human brain samples showed that our method 1) run 10 times faster than likelihood ratio test with logistic regression for any data size, keeping the accuracy of detecting three-way interactions at around 85% and 2) detected a large number of three-way gene interactions we were looking for. Figure 2 shows a typical example of the detected interactions with *p*-value of -8.91, where two genes are correlated with each other under two classes or anti-correlated under the other class. We confirmed the plausibility of this interaction in terms of the biological literature.

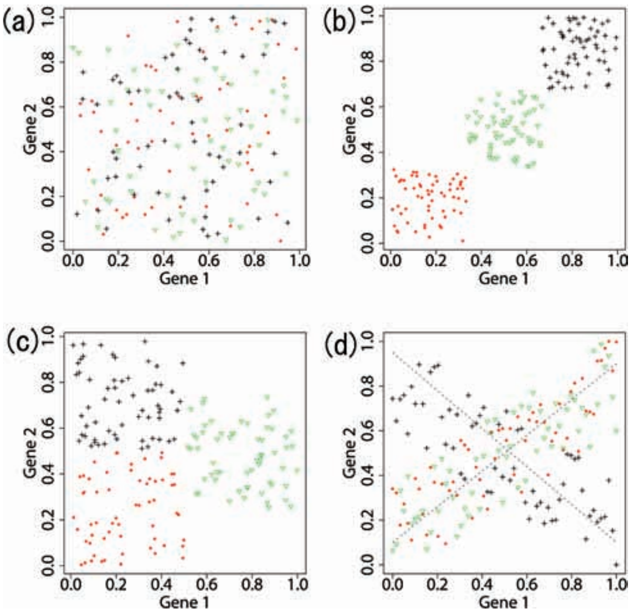

Figure 1.

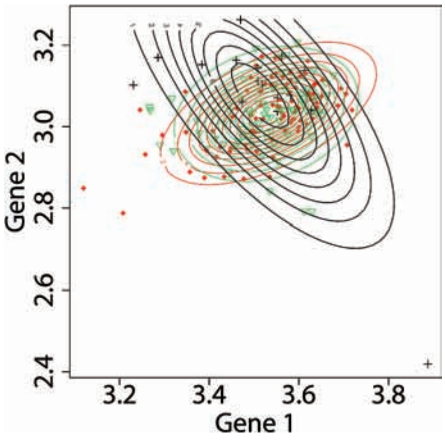

Figure 2.

2007–30 September 2010.

Takigawa I, Multifaceted Exploration of Nonhomogeneous and Ambiguous Data by Combining Partial Similarities, Grant-in-Aid for Young Scientist (B), 1 April 2008–31 March 2011.

Shiga M, Integrative Data Mining Based on Structure Analysis of Biological Networks, Grant-in-Aid for Young Scientist (B), 1 April 2008–31 March 2010.