Genome-wide three-way gene interactions from transcript and genotype data

Mitsunori Kayano\textsuperscript{1} Ichigaku Takigawa\textsuperscript{1} Motoki Shiga\textsuperscript{1} Koji Tsuda\textsuperscript{2} Hiroshi Mamitsuka\textsuperscript{1} 

\texttt{kayano@kuicr.kyoto-u.ac.jp takigawa@kuicr.kyoto-u.ac.jp shiga@kuicr.kyoto-u.ac.jp koji.tsuda@aist.go.jp mami@kuicr.kyoto-u.ac.jp}

\textsuperscript{1} Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
\textsuperscript{2} Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-42, Aomi, Koto-ku, Tokyo 1350064, Japan

Keywords: gene expression, genome-wide association study, logistic regression, SNP

Introduction

We address the issue of finding a three-way gene interaction, i.e. two interacting genes in expression under the genotypes of another gene, given a dataset in which expressions and genotypes are measured at once for each individual. This issue can be a general, switching mechanism in expression of two genes, being controlled by categories of another gene, and finding this type of interaction can be key to elucidating complex biological systems\textsuperscript{[1, 2]}. The most suitable method for this issue is likelihood ratio test using logistic regression, which we call interaction test, but a serious problem of this test is computational intractability at a genome-wide level.

Method

We developed a fast method for finding three-way gene interactions from transcript- and genotype-data. Our method prunes large part of input combinations, to which interaction test does not have to be applied. The following two cases are pruned in a main procedure of the proposed method: (a) expression values are randomly distributed in terms of classes, and (b) expression values can be easily categorized into three classes corresponding to the genotypes of a SNP. In Figure 1, panel (a) shows expression values being just randomly distributed; (b) shows expression values being easily categorized into three classes: +, • and △. We are not interested in (a) and (b) but in (c), which shows that the correlation in expression between two genes differs for each class. More concretely, two genes are positively correlated with each other for one class, whereas negatively correlated with each other for another.

![Figure 1: Three examples of distributions of expression values using simulated data sets.](image)

(a) randomly distributed (b) easily categorized into three classes (c) switching mechanism
The input combinations of (a) and (b) can be pruned by using randomness test and linear discriminant analysis (LDA), respectively. The randomness test is a combination of multivariate analysis of variance (MANOVA) and Box’s M test, and is here called by mean-covariance (MC) test. MANOVA and Box’s M test are statistical tests of homogeneities of means and covariances respectively, and MC test examines complete homogeneity with respect to means and covariances simultaneously. Only MC test can successfully detect (a) and is expected to work on real data. Our procedure is as follows: We first check each pair of a gene and a SNP by LDA. That is, if expression values of a gene are separable by genotypes of the SNP, this pair is pruned. We then check each combination of the remaining pair and another gene by LDA in a similar manner and then by MC test. In fact, if a combination is regarded as a random one, the combination is pruned. Finally interaction test is applied to the remaining combinations.

Results and Discussion

We first examined the improvement in time efficiency by our proposed method, which we call Fast finding Three-way Gene Interactions (FTGI) against the approach of running Interaction Test Only (ITO) over all possible combinations. Figure 2 shows the real computation time of ITO and FTGI, when we changed the number of combinations randomly chosen from a source dataset. The $\alpha_m$ in this figure is the significant level of MC test, and efficiency of FTGI can be affected by the value of $\alpha_m$. This figure clearly shows that as $\alpha_m$ decreased, the amount of running time of FTGI became smaller for any size of inputs, by pruning a larger number of them. In particular, at $\alpha_m$ of 0.001, FTGI runs approximately 10 times faster than ITO.

We further checked the pruning rate and the pruning accuracy. The pruning rate is the ratio of pruned combinations to all input combinations, and the pruning accuracy can be defined as the overlap between the resultant top 100 combinations by p-values of ITO and those of FTGI. If $\alpha_m$ is reduced to 0.001, 94% inputs can be pruned, keeping the pruning accuracy of 85%.

We applied our method to hundreds of millions of input combinations generated from a dataset on human brain samples [3] and detected three-way gene interactions with small p-values. To check the reliability of our results, we first conducted permutations by which we can show the obtained p-values are significantly smaller than those obtained from permuted null examples. We then used GEO (Gene Expression Omnibus) to generate gene expression datasets with binary classes to confirm the detected three-way interactions by using these datasets and interaction test. The result showed us some datasets with significantly small p-values. These results show the potential of our approach to explicate complex biological systems appearing in modern biology and medical sciences.

Acknowledgments: This work has been supported in part by BIRD of Japan Science and Technology (JST), JSPS KAKENHI (grant number: 20700134, 20700269, 21680025), the Functional RNA Project of New Energy and Industrial Technology Development Organization (NEDO).

References