Regular Paper

An algorithm for searching multiple SNP interactions

Naoto Ikeda[1] Hiroto Saigo[1]

Abstract: With the recent advance in analyzing biological sequence data, more and more information is becoming available for identifying the genes-disease relationship. For unveiling such a relationship, recent approaches often employ case-control study, where SNPs of a group of patients and a group of healthy people are collected first, and the difference of SNPs between the two groups are investigated subsequently. However, due to prohibitive amount of computation, most of existing study has focused on analysis of single SNP vs single phenotype relationship. In this study, we propose a method for searching multiple SNP interactions that affect on phenotype. Computational experiments on simulated data demonstrates effectiveness of our approach.

1. Introduction

With the advancement in technologies to analyze biological sequence data, more and more data are stored, and a development of a method for efficiently analyzing them is desired. Because of its cheap generating cost, SNP (Single Nucleotide Polymorphism) data is becoming popular for unveiling the disease-gene relationship. SNP is considered as genetic information that form differences between individuals. In a typical case-control study, SNPs of a group of patients (case) and a group of healthy people (control) are collected, and the difference in SNPs between the two groups are investigated [3]. There exists many research on this line, and we assume the same setting in this study. Most of the existing research has focused on a situation where only one SNP triggers the expression of the phenotype. This is simply because of the computational burden required for investigating multiple SNP interactions. For example, suppose we have a set of p SNPs, then considering k-way interaction requires considering p^k combinations. Since the number of SNPs p is typically in the order of hundreds of thousands, searching for k-way interaction is computational intensive for even a small k. Still, there are many studies that report the evidence of existence of more than two-way interactions related to the phenotype [4]. In this paper, we take a branch-and-bound approach under the assumption that k-way interaction occurs provided the existence of k-1 way interaction. For traversing huge space of combination, we employ an efficient method for generating candidates without overlapping [2], and bound the search space by a test based on AIC (Akaike Information Criterion) [1].

In computational experiments based on simulated data set, we demonstrate the effectiveness of our approach compared to the brute-force search which does not employ pruning but instead search entire space. We also demonstrate the robustness of our approach on simulated data with Gaussian noise added to phenotype.

[1] Kyushu Institute of Technology, 680-4 Kawazu 820-8502, Iizuka, Fukuoka, Japan

2. Algorithm

We employ branch-and-bound approach under the assumption that k-way interaction occurs provided the existence of k-1 way interaction. For efficiently traversing such space, we employ backtrack search method that generates candidate nodes without overlaps [2]. The resulting search space is displayed in Fig. 1. Another major part of branch-and-bound is the bounding part. For this purpose, we employ AIC (Akaike Information Criterion) for model selection [1]. Suppose we are at a node with SNPs x_1, x_2 in Figure 1, and have to decide either extend the search space to a node with SNPs x_1, x_2, x_3 or not. The decision is based on comparison of the following two models.

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 \quad (1)$$

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1 x_2, \quad (2)$$

where $y \in R^n$ denotes phenotype, $x \in \{0, 1\}$ denotes the binary indicator of an SNP, and α_i is a coefficient for each variable (α_0 for offset). If equation (1) fits better than equation (2) to given data in terms of some criterion, we can stop considering further interactions. If equation (2) fits better than equation (1) to given data, then further SNPs are considered for interaction. As a criterion for comparing two models, we employ AIC defined below.

$$AIC = -2 \log\left(\frac{RSS}{n}\right) + 2p. \quad (3)$$

where p is the number of variables in equation (1) or (2), and RSS (Residual Sum of Squares) is defined as:

$$RSS = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2, \quad (4)$$

where y is a true phenotype and $\hat{y}$ is predicted phenotype. In general as we use more variables, RSS and corresponding negative log-likelihood keep increases, but the second term in AIC penalizes adding too much variables. Indeed, this property is ideal in

our situation where we desire to minimize huge combinatorial search space as much as possible.

The effect of introducing pruning is illustrated in Figure 1 and Figure 2. Notice the shrinkage of the search space from Figure 1 to Figure 2. For example in Figure 1, there is no interaction observed by adding an SNP x_3 to SNPs $\{x_1, x_2\}$ in terms of AIC, so the search is stopped there, while the search must reach the leaf node if no pruning condition is employed (Figure 1).

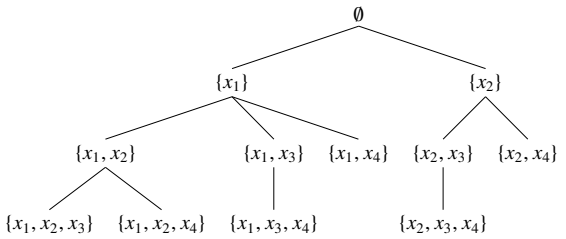

Fig. 1 A schematic figure of search space without pruning

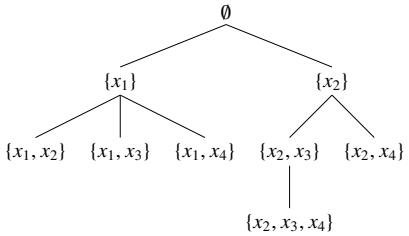

Fig. 2 A schematic figure of search space with pruning. Notice the shrinkage of the search space from Figure.1

3. Experiments

We simulated a data set consisting of 100 examples corresponding to 100 people, and added artificially generated 3-way interaction to the 15 of them. Phenotype are generated from normal distribution $N(0, 1)$, and the highest 15 values are assigned as labels to SNP data having 3-way interaction, and the labels of the rest (85) are randomly assigned to the rest of the SNP data. In order to verify the effectiveness of our method with pruning, our method was compared with the brute-force search which does not employ pruning at all. In the experiment, the number of SNPs were varied from 100 to 300.

In order to verify robustness of our method against noise in phenotype, we added Gaussian noise $N(0, \sigma^2)$ to the phenotype. The amount of σ^2 was changed for the range of 0.01, 0.1, 0.5 and 1.

4. Results

Figure 2 shows the time spent until completing the search as a function of the number of SNPs. It is observed that time spent for brute-force search method increases exponentially, while time spent for the proposed method was much smaller. Indeed, when the number of SNPs was 200, our approach was about 100 times faster.

Table 1 shows the robustness of our method against noise. It is observed that it could successfully find the true 3-way interaction until we add Gaussian noise $N(0, 0.5)$. However, it failed to find the true 3-way interaction when we added Gaussian noise $N(0, 1)$. Considering that the original phenotype are generated from $N(0, 1)$, our method has robustness to noise in phenotype.

5. Conclusion

We have proposed a novel algorithm for searching interactions among SNPs that has effect on phenotype. It implements efficient branch-and-bound search, and was demonstrated to run much faster than a brute-force search. Not only it runs fast, but also demonstrated was the robustness against noise in phenotype, which is a very likely setting in real data.

Our future work includes performing further experiments on simulated and real-world data to investigate the possibility and limitation of the method.

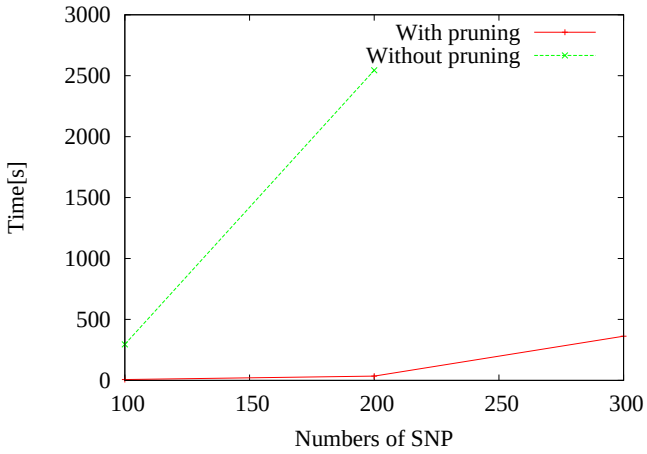

Fig. 3 Evolution of computation time while changing the number of SNPs

Table 1 Robustness of the proposed method against Gaussian noise

Variance(σ^2)	Time	Search space	True Interaction Found
0.01	6.89	58099	Yes
0.1	5.52	47576	Yes
0.5	7.95	66090	Yes
1	3.51	33285	No

References

[1] H.Akaike: "Information theory and an extension of the maximum likelihood principle", *2nd International Symposium on Information Theory*, Petrov, B. N., and Csaki, F. (eds.), Akadimiai Kiado, Budapest: 267-281 (1973).

[2] T.Uno, M.Kiyomi and H.Arimura: "LCM ver.3: Collaboration of Array, Bitmap and Prefix Tree for Frequent Itemset Mining", *ACM*,pp. 77-89 (2005)

[3] X.Zhang, F.Zou and W.Wang: "FastANOVA: an Efficient Algorithm for Genome-Wide Association Study", *ACM*, Vol. 3, pp. 821-829 (2008)

[4] J.Li, B.Horstman and Y.Chen: "Detecting epistatic effects in association studies at genomic level based on an ensemble approach",*Bioinformatics* 27(13), 222-229 (2011)

[5] J.Li and Y.Chen: "Generating samples for association studies based on HapMap data",*BMC Bioinformatics* 9, 44 (2008)