Application of Data Mining to Genetic Linkage Analysis

Nobutaka Mitsuhashi [1] mituhasi@ims.u-tokyo.ac.jp
Haretsugu Hishigaki [1 2] hisigaki@ims.u-tokyo.ac.jp
Toshihisa Takagi [1] takagi@ims.u-tokyo.ac.jp

[1] Human Genome Center,
Institute of Medical Science, The University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108 Japan

[2] Otsuka GEN Research Institute,
OTSUKA Pharmaceutical Co., Ltd.,
463-10 kagasuno Kawauchi-cho, Tokushima 771-01 Japan

Abstract

Genetic linkage analysis tools, which are used to determine the genetic distance and order among genetic markers, take too much computational cost. To improve the efficiency, we are investigating the applicability of data mining to linkage analysis from both statistical and practical aspects. In this paper, we report the results of the investigation and discuss the problems.

1 Introduction

One of the aims of genetic linkage analysis is to make a genetic map, that is, to determine the distance and order among genetic markers on a chromosome. The order is based on the genetic distance between two markers, which is defined as a recombination fraction calculated from the genotypes of the markers. To compute the order with the maximum likelihood, we have to calculate that of $n!/2$ possible orders for n markers. Approximate methods are used to reduce such computational cost. However, even those methods can not easily determine the order of more than ten markers at once.

To solve these problems, we tried to use one of data mining algorithms, Agrawal's association rule discovery algorithm[1], which finds the set of items that frequently occur together from a large number of tuples. It is so simple that only two simple measures *support* and *confidence* are used to discard uninteresting patterns. Concerning application to linkage analysis, the computational cost of the algorithm does not increase exponentially with respect to the number of the markers. If we can appropriately interpret mined patterns, it is possible to determine the order of multiple markers simultaneously. In addition, this method can be used to confirm the result of other linkage analysis tools, because the simplicity of the above two measures allows one to easily check whether the result of mining is qualitatively correct. Phenotypic and genotypic data can be uniformly treated in data mining as well. It may be useful for finding the relationships between phenotypes and genotypes.

2 Methods

We examine the statistical and practical applicability of Agrawal's algorithm, giving an example of F2 intercross mating system as illustrated in Figure 1. Genotypic data of an F2 generation (filial

generation no. 2) individual are given as a input in the form of a tuple, (... genotype of marker α,...,genotype of marker β,...). For F2 intercross, there are three genotypes: a homozygote for the alleles from the grandfather, a homozygote for the alleles from the grandmother, and a heterozygote for both types of alleles. In this paper, a capital letter means the allele from the grandfather, while a small letter from the grandmother. For example, A/a denotes that the genotype of marker α is heterozygous for both types of alleles. Besides, *support* s is defined for each genotypic pattern that means s % of all tuples contain such a genotypic pattern. For example, if there are one hundred tuples, twenty of which contain a genotype A/A and b/b for marker α and marker β, the support of this genotypic pattern is 20 %.

If the number of F2 individuals is great enough, the support of a particular genotypic pattern approximates the occurrence probability of it. The occurrence probability P with respect to two markers is a function of only one recombination fraction θ, which is monotone in the whole range of a recombination fraction ($0 \leq \theta \leq \frac{1}{2}$). Therefore we are able to conversely calculate a recombination fraction from the occurrence probability of a particular genotypic pattern. The occurrence probability P of genotypic pattern A/A and b/b for marker α and marker β is $\frac{1}{4}(1-\theta_{\alpha\beta})^2$. The recombination fraction $\theta_{\alpha\beta}$ is $1-2\sqrt{P}$ ($\frac{1}{16} \leq P \leq \frac{1}{4}$).

To investigate practical usefulness, we compared the genetic distance obtained from data mining with that from an existing linkage analysis tool, MAPMAKER/EXP, by using genotypic data containing 46 F2 individuals, and found that the result of data mining is qualitatively correct, even if the number of F2 individuals is comparatively small. Getting genotypic data of a large number of F2 individuals involves a lot of efforts in biological experiments. This result is crucial for practical use.

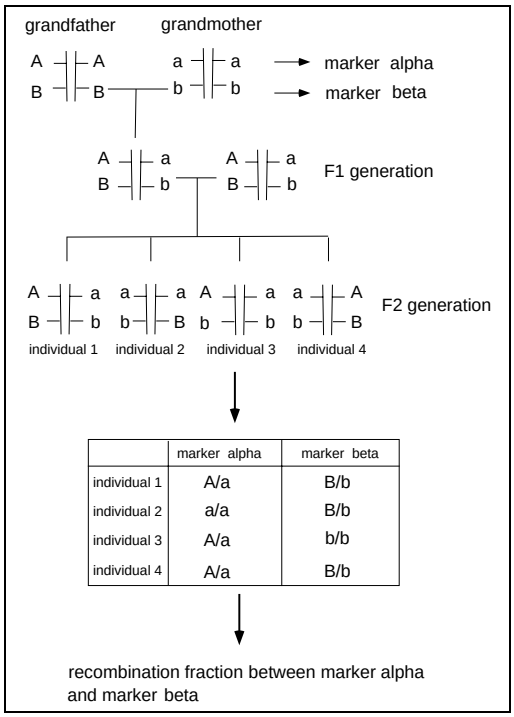

Figure 1. Data mining process in genetic linkage analysis

3 Future Works

We will establish the relevant interpretation of mined patterns for multi-point analysis, where genetic recombinations in multiple loci are considered simultaneously. The method mentioned above is not directly applicable, because the occurrence probability of a genotypic pattern is a function of two or more than two recombination fractions. In the future, we would like to develop this method to detect a disease gene locus.

Acknowledgment

This work was supported in part by a Grant-in-Aid (08283103) for Scientific Research on Priority Areas from The Ministry of Education, Science, Sports and Culture in Japan.

References

[1] Agrawal,R., Imielinski,T. and Swami,A., "*Mining Association Rules between Sets of Items in Large Databases*", ACM SIGMOD, pp.207–216, 1993.