Applying an Association Rule Discovery Algorithm to Multipoint Linkage Analysis

Nobutaka Mitsuhashi[1] mituhasi@ims.u-tokyo.ac.jp
Haretsugu Hishigaki[2] hisigaki@ims.u-tokyo.ac.jp
Toshihisa Takagi takagi@ims.u-tokyo.ac.jp
Human Genome Center, Institute of Medical Science, The University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108 Japan

Abstract

Knowledge discovery in large databases (KDD) is being performed in several application domains, for example, the analysis of sales data, and is expected to be applied to other domains. We propose a KDD approach to multipoint linkage analysis, which is a way of ordering loci on a chromosome. Strict multipoint linkage analysis based on maximum likelihood estimation is a computationally tough problem. So far various kinds of approximate methods have been implemented. Our method based on the discovery of association between genetic recombinations is so different from others that it is useful to recheck the result of them. In this paper, we describe how to apply the framework of association rule discovery to linkage analysis, and also discuss that filtering input data and interpretation of discovered rules after data mining are practically important as well as data mining process itself.

1 Introduction

To detect a disease gene locus on a chromosome, genetic linkage analysis technologies have been developed. They consist of two phases: (1) making a genetic map, that is, determining the distance and order among genetic marker loci, and (2) locating a disease locus in a genetic map by discovering associations between the marker loci and the phenotypic traits of a disease. In this paper, we will address the first phase using an association rule discovery algorithm.

How is the distance between marker loci calculated? Maximum likelihood estimation has been widely used, which is a statistical method of estimating the values of parameters by maximizing the occurrence probability of samples. Genetic linkage analysis is formalized in terms of maximum likelihood estimation as follows. Sample data are genotypes of genetic markers per individual. Parameters to be estimated are recombination fractions between two loci, each of which is approximately equivalent to the genetic distance between loci. By maximizing the likelihood of samples, recombination fractions between two loci are calculated.

To determine the order of loci, we have to consider more than three loci. Certainly, we can determine the order by pairwise comparisons of recombination fractions between all available loci. However, the accuracy of linkage analysis depends upon how much information we can exploit from sample data. In the case of ordering loci, considering many loci simultaneously makes it possible to obtain much more information than considering only two loci [8]. If we could not, we would have to collect much more genotypes. A way of linkage analysis like this is called multipoint analysis, while two-point analysis is such that only two loci are considered simultaneously.

Strict multipoint analysis involves the calculation of the likelihood of possible orders and the choice of the best order among them. However, it is obvious that the calculation of the likelihood of possible

[1]Present affiliation and address are Bioscience Systems Department, Mitsui Knowledge Industry Co., Ltd., and 2-7-4 Higashi-Nakano, Nakano-ku, Tokyo 164 Japan.

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

orders is practically intractable for more than ten marker loci. For n loci, there are $\frac{n!}{2}$ possible orders. Finding an order with the highest likelihood requires $\frac{n!}{2}$ calculations of the likelihood. Multipoint linkage analysis is computationally hard.

From a viewpoint of efficiency, establishing an approximation method is crucial in multipoint linkage. For example, `order` command of MAPMAKER/EXP [6, 7] finds the best order as follows. First, the order of a subset chosen at random is determined by exhaustive $\frac{n!}{2}$ calculations of the likelihood. This initial order must be much more likely than the second best one. Next, the rest of loci are inserted into an appropriate interval of the initial order one by one.

No one can say that the order of loci obtained by these ordering methods is true on a actual chromosome, unless physical mapping is performed. Therefore, in order to confirm the accuracy of the result, various kinds of ordering criteria are required. Certainly, criterion mentioned above intuitively seems to be correct, however, they are the aggregation over intervals. So one can not logically argue whether such a criterion is always valid or sometimes concludes a wrong answer. On the other hand, our method, by which overlapping intervals are found from an association of genetic recombinations, is not only different in principle, but also it is understandable why it concludes a right answer.

Another reason why we focus on an association rule is that the simplicity of the framework makes it possible to handle different kinds of data uniformly. In [11], the authors discovered association rules between three kinds of heterogeneous data, amino acid sequences, protein structures and functions. In trait analysis, the second phase of linkage analysis mentioned in the beginning of this section, we have to handle heterogeneous data, genotypes and phenotypes [5].

KDD process roughly consists of three components: data mining, preprocessing input data and rule interpretation. The main component is data mining. In section 2, we explain how to find an association rule. Preprocessing requires a deep understanding of an application domain, that is, to understand a mating system and the characteristics of genotypic data. We touch on it in section 3. How to order three loci and more than three by the interpretation of discovered rules is described in section 4 and 5. In section 4, we discuss how to order three loci, so-called three-point analysis, and how to order more than three loci using the result of three loci in section 5. In section 6, we discuss the problem caused by a double recombinant and the solution to it. Every component is important to obtain desired knowledge. We are repeating the inspection of the final result and the improvement of each process.

2 Association Rule

Finding associations from a large amount of data efficiently is required in many application domains. A framework of an association rule and an efficient algorithm for finding such rules were presented by Agrawal et al. [1, 2]. They also applied association rules to the analysis of sales data. In this paper, we only apply their algorithm to an ordering problem. We won't change their framework at all.

We describe an overview of association rule discovery using an example of sales data. Input data are provided in the form of a table, that is, a set of tuples. For example, the first tuple of table 1 means that an eraser and a notebook are bought together in transaction 1. After mining, we obtain a set of association rules in the form of $X \Rightarrow Y$, where X and Y are items. Intuitive meaning of an association rule $X \Rightarrow Y$ is that when X holds, Y also holds. For example, an association rule $pencil = 1 \Rightarrow eraser = 1$ means that customers who buy a pencil also buy an eraser.

Association rules generated by a mining algorithm must be interesting to us. Agrawal used two measures to evaluate the interestingness of rules. *Support* of items X and Y ($sup(X \wedge Y)$) or a rule $X \Rightarrow Y$ ($sup(X \Rightarrow Y)$) is defined as the rate of tuples including both items. *Confidence* of a rule $X \Rightarrow Y$ ($conf(X \Rightarrow Y)$) is defined as the ratio of $sup(X \wedge Y)$ to $sup(X)$. Association rules with higher support and confidence than *minimum support* and *minimum confidence* given by user are considered significant. Assuming that minimum support and minimum confidence are 50 % and 75%, only one association rule $pencil = 1 \Rightarrow notebook = 1$ is discovered from the sales data in table 1.

trans_id	pencil	eraser	notebook
1	0	1	1
2	1	0	1
3	0	0	1
4	1	0	1

Table 1: Example of sales data

In our case, we can represent input data as table 2. The number of the rows and columns corresponds to that of individuals, from which genotypic data are collected, and that of loci to be ordered, respectively. A cell G_{i-j} of the table shows the genotype of locus j of individual i. How about the values of the genotype is described in the next section.

individual_id	G_{L1}	G_{L2}	G_{L3}
1	G_{1-L1}	G_{1-L2}	G_{1-L3}
2	G_{2-L1}	G_{2-L2}	G_{2-L3}
3	G_{3-L1}	G_{3-L2}	G_{3-L3}

Table 2: Table of input data

3 Preprocessing Input Data

The key to estimate the distance between marker loci is to know how often recombinations occurred between them. Given the genotypes of individuals as input data, how can we know about a recombination event? We have to explain a mating system, F2 backcross [10], from which genotypic data are collected.

F2 backcross is a kind of mating systems as illustrated in Figure 1. In F1 generation, a doubly heterozygous individual(AB/ab) is married to a doubly homozygous one (ab/ab) who has the same genotype as one of the grandparents. As a result, there are two kinds of genotypes of F2 individuals, heterozygote for grandpaternal and grandmaternal alleles (h) and homozygote for grandmaternal alleles (H). Genotypic data of F2 individuals are used in linkage analysis.

The most important property of F2 backcross is phase-known. One can uniquely find out whether a genetic recombination occurred or not. Figure 1 shows that nonrecombination on a paternal chromosome occurred in individuals 1 (AB/ab) and 2 (ab/ab) whose genotypes of loci 1 and 2 are identical. On the other hand, one recombination occurred in individuals 3 (Ab/ab) and 4 (aB/ab) whose genotypes are different.

Strictly, the fact that the genotypes of two loci are different means an odd number of recombinations. However, as far as concerning a chromosomal region less than about 100 cM, we don't have to take care. The assumption that a genetic recombination occurs on a chromosome under study at most once is called complete interference.

Considering the properties of F2 backcross, we can transform the original tuple (G_{L_i} representation) into an explicit representation of a genetic recombination between two marker loci (R_{L_i,L_j} representation) as follows:

$$\begin{array}{lll} \text{original tuple } (G_{L_i} \text{ representation}) & (G_{L1}, G_{L2}, G_{L3}, \cdots, G_{Ln-1}, G_{Ln}) & = \quad (H, h, H, \cdots, H, h) \\ & \Downarrow & \end{array}$$

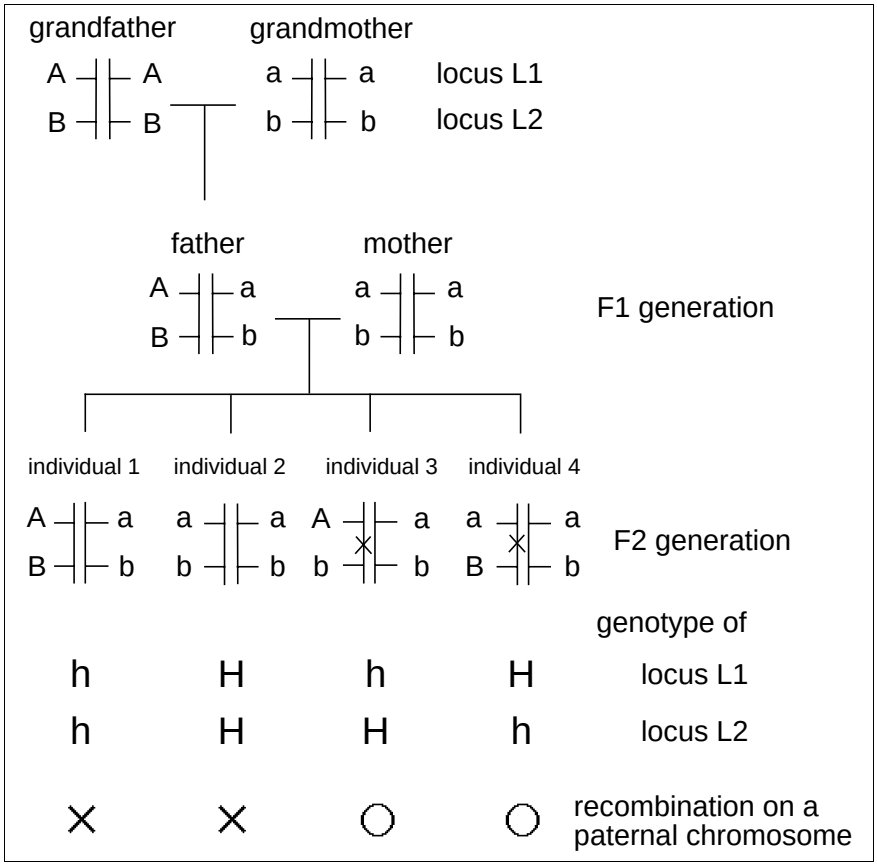

Figure 1: F2 backcross pedigree. 'X' between loci 1 and 2 shows one genetic recombination between them. In this paper, genotype 'h' denotes heterozygote for grandpaternal and grandmaternal alleles, while genotype 'H' homozygote for grandmaternal alleles.

original tuples

individual_id	G_{L1}	G_{L2}	G_{L3}
1	H	h	H
2	h	h	H
3	H	H	H

$\Downarrow$ data transformation of F2 backcross data

transformed tuples

individual_id	R_{L1L2}	R_{L1L2}	R_{L1L3}
1	1	1	0
2	0	1	1
3	0	0	0

Figure 2: Data transformation

$$\text{transformed tuple } (R_{L_i,L_j} \text{ representation}) \quad \underbrace{(R_{L1L2}, R_{L1L3}, \cdots, R_{Ln-1Ln})}_{\frac{1}{2}n(n-1) \text{ elements}} \quad = \quad (1, 0, \cdots, 1)$$

where $G_{L_i} = H$ and $G_{L_i} = h$ means that the genotype of locus i is heterozygous and homozygous, respectively, and $R_{L_iL_j} = 1$ indicates that a genetic recombination between loci i and j occurred, while $R_{L_iL_j} = 0$ corresponds to nonrecombination. For example, $G_{L1} = H$ and $G_{L2} = h$ implies $R_{L1L2} = 1$. Figure 2 illustrates the data transformation from the original tuples of F2 backcross data.

Association rules constituted by $R_{L_iL_j}$ items are easy to understand, when we interpret the rules. In the next section, we use association rules made from $R_{L_iL_j}$ items. However, original tuples that G_{L_i} items constitute are used as inputs of an association rule discovery program, due to time complexity. Details are explained in the following section.

4 Rule Interpretation for Three-Point Analysis

After data mining, we determine a locus order by giving relevant interpretation to discovered rules. In the beginning, we try to order three loci as the first step to ordering more than three. For three loci L1, L2, and L3, there are three intervals L1L2, L2L3, and L1L3. We focus on an association rule in the form of $R_{L1L2} = 1 \Rightarrow R_{L1L3} = 0$ under the assumption of complete interference. Intuitive meaning of the rule is that when a recombination for L1L2 occurred, that for L2L3 never occurred. This means that the locus order is L1-L2-L3. We will explain these things in terms of *confidence* rather than *support*.

More precise interpretation is given by finding a mapping from $conf(R_{L1L2} = 1 \Rightarrow R_{L1L3} = 0)$ to the order of three loci. The relationships between the order of three loci and the confidence is illustrated in Figure 3. The confidence ranges from 0 to 1, according to three patterns of the locus orders. We try to find a mapping from the confidence to the corresponding locus order as the inverse mapping or contraposition, after finding a mapping from each locus order to its confidence.

In case 1, that is, locus order L1-L2-L3, the above association rule intuitively means that whenever a recombination for L1L2 occurred, that for L2L3 never occurred. Interval L1L2 does not include interval L2L3. The confidence equals 1. In case 2, L1-L3-L2, $R_{L2L3} = 0$ does not always hold, when $R_{L1L2} = 1$ holds. The confidence ranges from 0 to 1. In case 3, L3-L1-L2, interval L1L2 overlaps with interval L2L3. The confidence of the rule is always 0, because a recombination for L1L2 always indicates that for L2L3.

Without unknown genotypes, solid arrows contain all mappings. Suppose that unknown genotypes may be included, however, another mappings shown by broken arrows have to be considered. For example, suppose that there are some tuples including $R_{L1L2} = 1$ and $R_{L2L3} = unknown$. The confidence may be less than 1, even if the order is L1-L2-L3. $R_{L2L3} = unknown$ reduces $sup(R_{L1L2} = 1 \wedge R_{L2L3} = 0)$ only, but it does not $sup(R_{L1L2} = 1)$. By the way, in the case of multipoint linkage analysis based on maximum likelihood, a numerical algorithm, EM algorithm [3], is frequently adopted to deal with missing data. Ours is a logically correct way to deal with missing data.

From figure 3, the contraposition that holds despite including unknown genotypic data is that $conf(R_{L1L2} = 1 \Rightarrow R_{L2L3} = 0) \neq 0$ implies that the order of three marker loci is either L1-L2-L3 or L1-L3-L2, not L2-L1-L3. In other words, locus L1 is not in the middle of three loci. To prove L1-L2-L3, the followings have to be satisfied:

$$conf(R_{L1L2} = 1 \Rightarrow R_{L2L3} = 0) \neq 0 \quad \Rightarrow \quad \text{Locus L1 is not in the middle of three loci.} \qquad (1)$$

$$conf(R_{L2L3} = 1 \Rightarrow R_{L1L2} = 0) \neq 0 \quad \Rightarrow \quad \text{Locus L3 is not in the middle of three loci.} \qquad (2)$$

At the same time, $conf(R_{L2L3} = 1 \Rightarrow R_{L1L3} = 0) = 0$ holds. However, this is not always true under no assumption of complete interference. As mentioned above, G_{L_i} representation is superior to $R_{L_iL_j}$ one with respect to time complexity. Time complexity is $O(n^3)$ in the case of G_{L_i}, while $O(n^4)$ in the

case of $R_{L_iL_j}$. Assume that n is the number of loci. In the case of $R_{L_iL_j}$ representation, $\frac{1}{2}n(n-1)$ items are required for n loci. Two itemsets are calculated to make association rules like (1) and (2), so the number of association rules is $\{\frac{1}{2}n(n-1)\}^2$, that is, $O(n^4)$. In the case of G_{L_i} representation, only $2n$ items are required for n loci, because $G_{L_i} = H$ and $G_{L_i} = h$ for one loci. Three itemsets are combined to make association rules like (3) and (4), so the number of generated rules is $(2n)^3$, that is, $O(n^3)$. Time complexity of three-point analysis is $O(n^3)$.

Formulas (1) and (2) are translated into (3) and (4), respectively. three-point analysis is performed by using these formulas instead of (1) and (2).

$$sup(G_{L1} = H \wedge G_{L2} = h \Rightarrow G_{L3} = h) \quad + \quad sup(G_{L1} = h \wedge G_{L2} = H \Rightarrow G_{L3} = H) \neq 0 \qquad (3)$$

$$sup(R_{L2} = H \wedge R_{L3} = h \Rightarrow G_{L1} = h) \quad + \quad sup(R_{L2} = H \wedge G_{L3} = h \Rightarrow G_{L1} = h) \neq 0 \qquad (4)$$

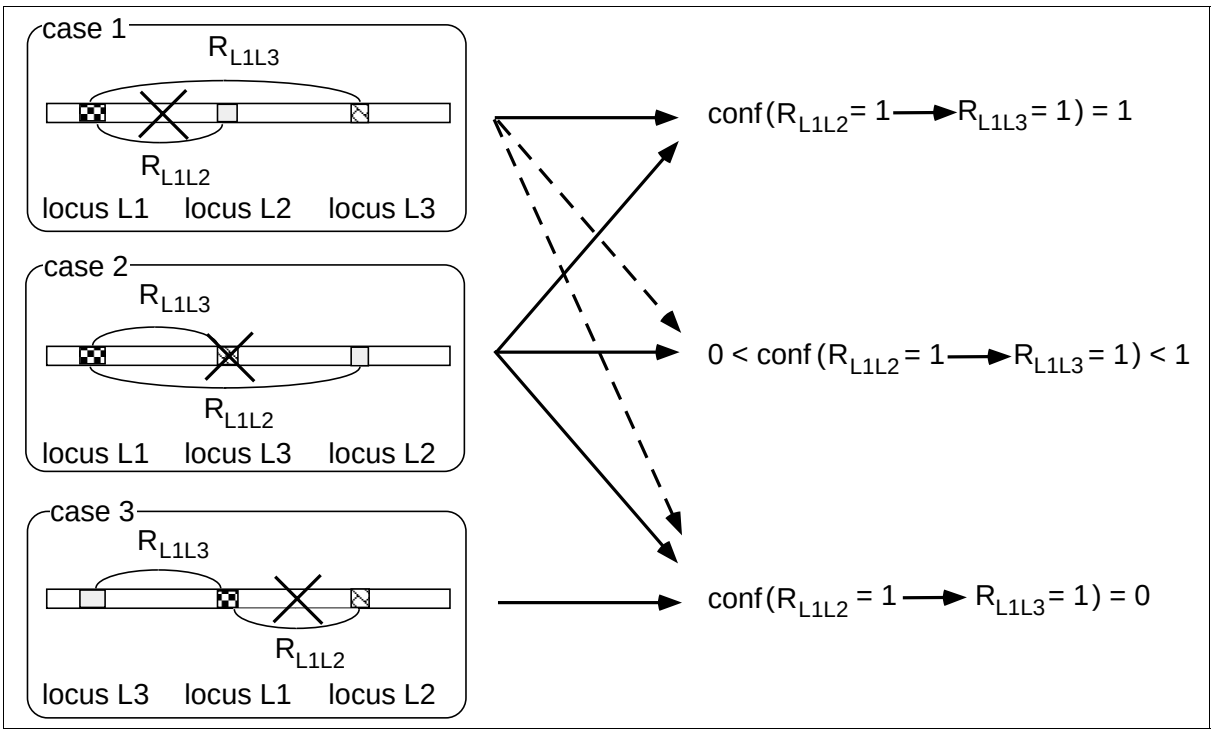

Figure 3: Relationships between the order of three loci, L1, L2 and L3, and the confidence of an association rule $R_{L1L2} = 1 \Rightarrow R_{L1L3} = 1$. A solid arrow shows a mapping from a particular order of three marker loci to the confidence of the rule without unknown genotypic data. Both solid and broken arrow show mappings, when unknown genotypic data are also included.

5 Ordering More Than Three Marker Loci

Three-point results, that is, the orders of all triplet marker loci, are used to order more than three. More than three loci are ordered using a divide-and-conquer technique. First, a pivot locus is chosen from a set of loci. A set of loci is divided into two subgroups separated by the pivot, according to three-point results including the pivot.

In figure 4, locus L3 is chosen as the first pivot. From two three-point results, L1-L3-L4 and L2-L3-L4, all loci are divided into two subgroups, group 1 and 2. Group 1 consists of loci L1 and L2,

while group 2 locus L4. The pivot L3 is added to both subgroups. In stead of a three-point result L2-L3-L4, L3-L1-L2 or L3-L2-L1 also supports that loci L1 and L2 belong to the same subgroup.

Group 1 should be further divided into two subgroups. Locus L2 is chosen as a pivot. According to a three-point result L1-L2-L3, group 1 is divided into group 1-1 and group 1-2, the member of which is locus L1 and L3, respectively. None of the groups are not divided, because all groups contain only two marker loci including a pivot. In a conquer stage, the order of the subgroups are determined by the previous pivot. We can find that group 1-2 is adjacent to group 2, because group 1-2 contains a pivot locus L3. The order of the subgroups is group 1-1 - group 1-2 - group 3. Finally, the order of four loci turns out to be L1-L2-L3-L4. The average number of the division is $O(\log n)$, where n is the number of loci.

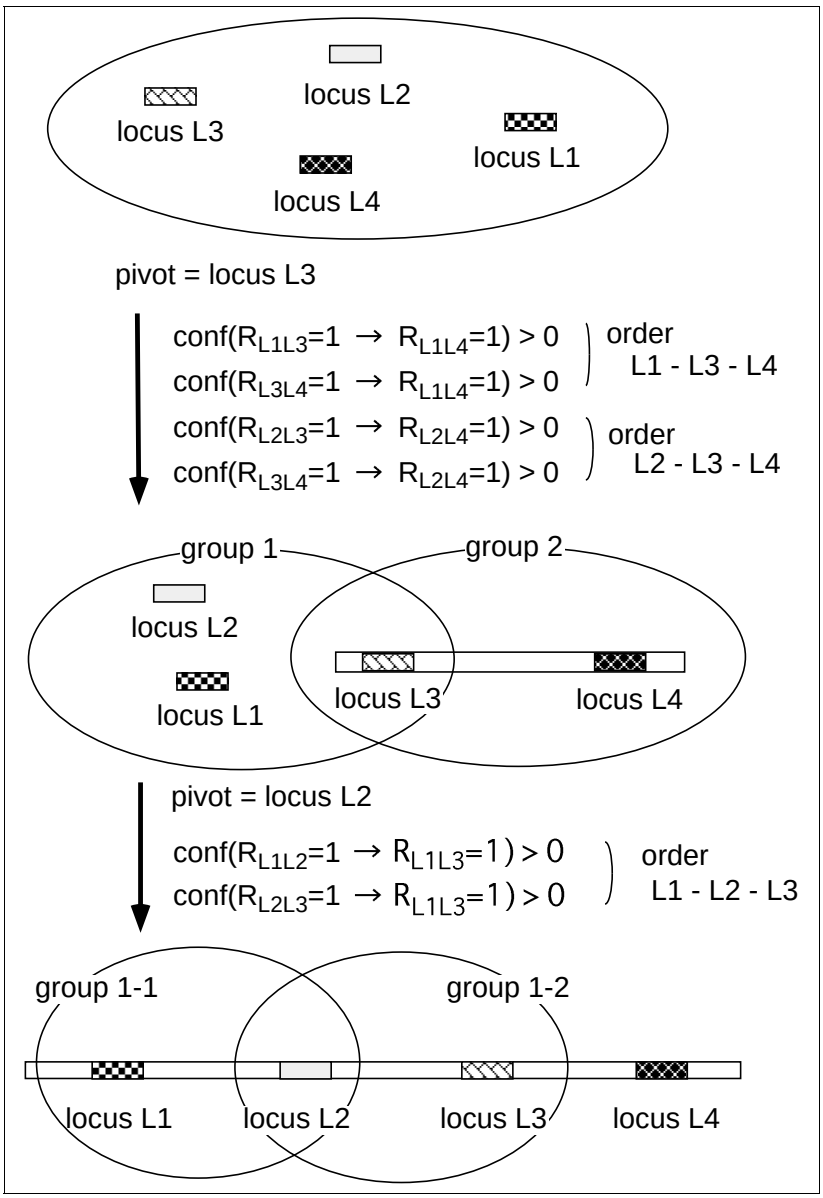

Figure 4: Ordering more than three loci using a divide-and-conquer technique.

6 Problem of a Double Recombinant

In the previous section, we gave the interpretation of the discovered rules under complete interference. Our argument is not valid, however, if we choose triplet loci including a double recombinant illustrated in figure 5. We have not yet established a way to directly determine the order from such data. If we can exclude such a triplet including a double recombinant, we can determine the order under complete interference. In this section, we describe how to exclude such triplets. Concerning three loci, we show the possible patterns of recombinants in table 3. In the following paragraphs, we assume that a true order is L1-L2-L3.

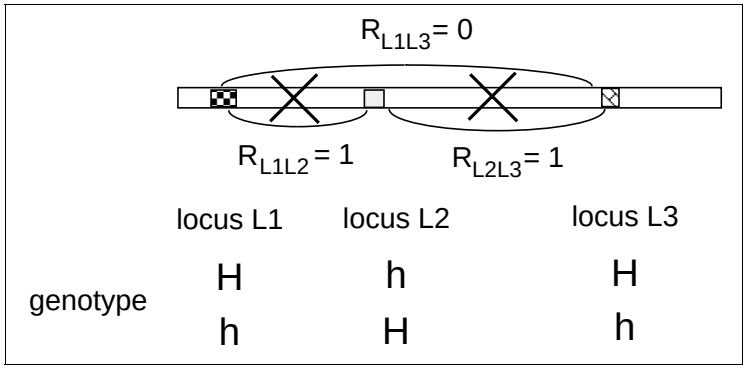

Figure 5: Double recombinant between three loci

pattern	R_{L1L2}	R_{L2L3}	R_{L1L3}	L1-L2-L3	L1-L3-L2	L2-L1-L3
P1	0	0	0	0	0	0
P2	1	0	1	1	1	2
P3	0	1	1	1	2	1
P4	1	1	0	2	1	1

Table 3: Possible recombination patterns of individuals. The columns from the second to the fourth show the recombination patterns of individuals. The rightmost three columns show the number of recombinations under the possible orders of three loci. In the case of P2, a recombination occurs in interval L1-L2 and L1-L3. Assuming that the order is L1-L2-L3, only one recombination occurs between L1 and L2. Two recombinations occur simultaneously in the case of L2-L1-L3.

When all four kinds of individuals, P1, P2, P3 and P4, are found in genotypic data, two necessary conditions (1) and (2) for proving that one of three loci is not in the middle of three hold for all three loci. In this case, there is no evidence to confirm which order is true among possible three. Such a triplet does not contribute to ordering, but does not lead to a wrong answer, because we can easily distinguish this pattern. This kind of triplets are ignored.

A worse case is that a double recombinant P4 is included, but both a single recombinant P2 and P3 are not included. For example, we observed only three patterns, P1, P3 and P4. Under complete interference, it is concluded that L2-L1-L3 is a true order. Such a case must be excluded from consideration, but it is not easily to distinguish them.

In what situation such a case is observed? It is a case that the length of interval L1-L2 is much shorter than that of interval L2-L3. In this case, the number of P2 sometimes may fall into zero, while that of P4 is not zero. A recombination of interval L1-L2 hardly occurs, compared with that

of L2-L3. Therefore, whenever a recombination of interval L1-L2 occurs, that of interval L2-L3 also occurs. In terms of *support*, $sup(R_{L1L2} = 1)$ is much smaller than $sup(R_{L2L3} = 1)$. The probability that at least one P2 is observed among L1-L2 recombinants is $P(P2|R_{L1L2} = 1) = 1 - (sup(R_{L2L3} = 1))^N$, where N is the number of L1-L2 recombinants, that is, $sup(R_{L1L2} = 1)\times$ (total number of individuals). In the worst case, when $sup(R_{L1L2} = 1)$ and N are close to $\frac{1}{2}$ and 1, respectively, the probability $P(P2|R_{L1L2} = 1)$ could be down to $\frac{1}{2}$. We have to check whether $P(P2|R_{L1L2} = 1)$ and $P(P3|R_{L2L3} = 1)$ are great enough to confirm the result of three-point analysis, L1-L2-L3. This is done by postprocessing.

7 Experimental Results

We applied our method to F2 backcross data of chromosome 1 of BSB mouse, which is available from The Jackson Laboratory[1]. Genetic maps per chromosome are also available as well as genotypic data. Chromosome 1 of BSB mouse contains genotypes of 33 genetic markers of 94 F2 individuals. The genetic distance of chromosome 1 is approximately 120 cM. The genetic map shows that two recombinations occurred between distant loci in some individuals. We used an association rule discovery algorithm, *apriori* [2], implemented in C. Both preprocessing and postprocessing are performed with several perl scripts.

All results of triplets not including double recombinants are consistent with the results of `compare` command of MAPMAKER/EXP, which executes $\frac{n!}{2}$ calculations of the likelihood. The data from The Jackson Laboratory also include a double recombinant as mentioned in the previous section. For example, $sup(R_{D1MitA1\_D1Mit112} = 1) = \frac{52}{94}$, is thirteen times greater than $sup(R_{D1Mit112\_D1Mit150} = 1) = \frac{4}{94}$. A wrong order D1MitA1 - D1Mit150 - D1Mit112 is derived without postprocessing, whereas a correct order is D1MitA1 - D1Mit112 - D1Mit150. However, the postprocessing excludes such a triplet.

8 Related Works

There are other approximate criteria based on two-point analysis. SAL(the sum of adjacent log-likelihoods) says that the best order is the order with the maximum sum of adjacent log-likelihoods. SAR (the sum of adjacent recombination fractions), SARF [12] and MDMAP [4] say that the best order is that with the smallest sum of adjacent recombination fraction. These criteria are used to make a preliminary map, because the accuracy is not guaranteed.

CPROP [9] is also a rule-based approach for constructing genetic maps as well as ours. The input data are a set of partial orders of loci, not genotypes. It is not a multipoint analysis program, but can deal with several kinds of partial orders and constraints that may be inconsistent with each other. Therefore, the calculation is so complicated that the time complexity is $O(n^5)$, where n is the number of loci.

9 Conclusions

In this paper, we proposed a novel and intuitively understandable approach to multipoint linkage analysis using an association rule discovery algorithm. An idea of finding association between genetic recombinations is completely different from maximum likelihood estimation that most multipoint linkage analysis programs adopt. We also presented some heuristics in order to deal with genotypic data, a part of which are missed or include double recombinants.

[1] http://lena.jax.org/resources/documents/cmdata

Acknowledgments

This work was supported in part by Grant-in-Aid for Scientific Research on Priority Areas, "Genome Science" from The Ministry of Education, Science, Sports and Culture in Japan. We also thank Dr. Noguchi and Dr. Satou for providing us with apriori program implemented in C, and Dr. Morishita for his advice on improvement of time complexity of three-point analysis.

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.

[2] Agrawal, R., Srikant, R., "Fast Algorithms for Mining Association Rules", *Proc. of the 20th Int'l Conference on Very Large Databases*, pp.487- 499, 1994.

[3] Dempster et al., "Maximum likelihood from incomplete data via the EM algorithm", *J. R. Statist. Soc.*, Ser. 39, pp.1–38, 1977.

[4] Falks, C. T., "A simple scheme for preliminary ordering of multiple loci: Application to 45 CF families. In Multipoint mapping and linkage based upon affected pedigree members", *Genetic Analysis Workshop 6*, pp. 17–22, 1989.

[5] Hishigaki, H. et al., "Prediction of a Disease Gene Locus by a Data Mining Algorithm", *HGM'97*, Mar. 1997(Toronto).

[6] Lander, E. S., and Green, P., "Construction of multilocus genetic linkage maps in humans", *Proc. Natl. Acad. Sci. USA*, Vol. 84, pp. 2363–2367, Apr. 1987.

[7] Lincoln, S. E., Daly, M. J., and Lander, E. S., "Constructing Genetic Linkage Maps with MAPMAKER/EXP version 3.0: A Tutorial and Reference Manual", *A Whitehead Institute for Biomedical Research Technical Report*, Jan. 1993.

[8] Lathrop, G. M. et al., "Strategies for multilocus linkage analysis in humans", *Proc. Natl. Acad. Sci. USA*, Vol. 81, pp. 3443–3446, Jun. 1984.

[9] Letovsky, S., and Berlyn, M. B., "CPROP: A Rule-Based Program for Constructing Genetic Maps.", *Genomics*, Vol. 12, pp. 435–446, 1988.

[10] Ott, J., *Analysis of human genetic linkage*, Johns Hopkins University Press, 1991.

[11] Satou, K. et al., "Finding Association Rules on Heterogeneous Genome Data", *Proc. of the Pacific Symposium on Biocomputing '97 (PSB'97)*, pp.397-408, Jan. 1997(Hawaii).

[12] Weeks, D. E., "New mathematical methods for human gene mapping", Ph.D. diss., University of California, 1988.