Tracing Synergetic Behavior of the QTLs Affecting Oral Glucose Tolerance in the OLETF Rat

Akihiro Nakaya [1] **Haretugu Hishigaki** [2] **Shinichi Morishita** [1]

nakaya@ims.u-tokyo.ac.jp hisigaki@ims.u-tokyo.ac.jp moris@ims.u-tokyo.ac.jp

[1] Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

[2] Otsuka GEN Research Institute, Otsuka Pharmaceutical Co., Ltd., 463-10 Kagasuno, Kawauchi-cho, Tokushima 771-0192, Japan

Abstract

The synergetic effects of multiple marker loci regarding quantitative traits such as blood glucose level have attracted interest. In the OLETF model rat of non-insulin dependent diabetes mellitus (NIDDM), our previous study focusing on the effects of multiple genetic factors has found significant marker combinations with respect to oral glucose tolerance (OGT) at 60 minutes after oral administration. Besides the interaction among markers at a particular time point, their correlated behavior in a time series is another interest. Based on the previous results, in this paper, we report the behavior of markers in a time series by using a series of measurements of OGT.

1 Introduction

Quantitative Trait Loci Analysis. Oral glucose tolerance (OGT), as well as factors such as body weight, fat weight, and insulin resistance, is an important quantitative trait significant to non-insulin dependent diabetes mellitus (NIDDM). Oral glucose tolerance (measured as the postprandial blood glucose level) is considered to be regulated by multiple *quantitative trait loci* (QTLs). In the investigation of these trait-causing loci, model rat strains of NIDDM have been developed, and some OGT-related loci have been mapped on the genome [3, 4]. In QTL analysis, the genotypes at marker loci and observation of quantitative trait value in the individuals are given as input data.

Linear Regression and LOD Score. The *interval mapping* method [7] based on a simple regression model has been widely used to map QTLs and found the OGT-related loci [2, 4]. The existence of a QTL within an interval flanked by a pair of neighboring marker loci is estimated along the genome. The logarithm of the likelihood ratio of the linkage between a marker locus and the quantitative trait against no linkage is called the *LOD score* and is calculated at each marker locus (or a putative locus in an interval with the genotype estimated from the flanking marker loci).

The OLETF Model Rat. A previous study employing the LOD score has identified OGT-causing QTLs on chromosomes in an F_2 intercross progeny of the Otsuka Long-Evans Tokushima Fatty (OLETF) rat [5]. The OLETF rat strain is an animal model of non-insulin dependent diabetes mellitus (NIDDM). These rats exhibit hyperglycemia, hyperinsulinemia, insulin resistance, and obesity, as well as showing glucose intolerance [5]. In our study we used F344 rats as a non-diabetic control strain. A cohort of male (female OLETF × male F344)F_2 intercross progeny including 157 rats was studied.

We used 279 microsatellite markers to determine the genotype of each individual. As shown in Fig. 1, in the F_2 intercross progeny, marker loci on the autosomes indicate the genotypes of the OLETF homozygote, the F344 homozygote, and the

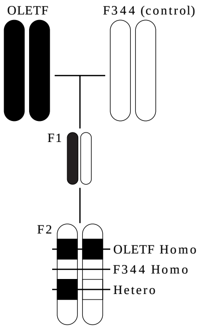

Figure 1: F_2.

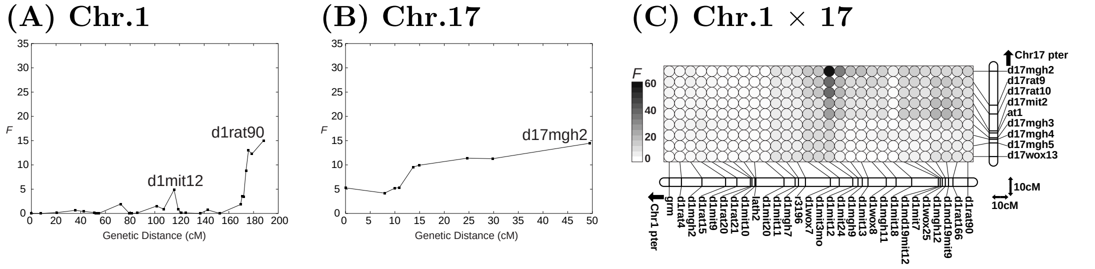

Figure 2: Effects of multiple markers. **(A, B)** Pointwise estimation of markers' significance along chromosomes 1 and 17. **(C)** Significance of marker pairs on chromosomes 1 × 17 (the black spot indicates a significant pair). We can observe the correlated effects between these chromosomes.

heterozygote (O/O, F/F, and O/F, respectively). For the marker loci on the X-chromosome, we also use the notations of O/O and F/F for the hemizygote genotypes of O and F. Recently, more than five thousand microsatellite markers have been identified and shown to be densely spaced throughout the entire genome [12]. Based on this fact, we assume that QTLs are linked to marker loci.

Multiple Factors. Pointwise estimation of the evidence for a QTL along the genome assumes that the markers are not correlated with each other. Therefore, the next task is to clarify the interactions between the trait-causing marker loci. In order to reflect the effects of the multiple marker loci on the explained trait value, a multiple linear regression model provides a theoretical expansion of the simple regression model. However, as mentioned by Zeng [13], a multiple linear regression model still assumes additivity of the QTL effects between loci. Even if one can construct a multiple linear regression model which adequately explains a quantitative trait, it is difficult to interpret the model (i.e., its partial coefficiencies) except when the markers behave independently of each other.

Actually, selectivity and correlation between marker loci have been found with respect to the glucose level at 60 min after oral administration in our previous study [10]. For example, Fig. 2A and Fig. 2B show the pointwise estimation of the significance of the O/O genotype at marker loci along chromosome 1 and 17. The significance is expressed by the F ratio (as discussed later, this quantity is equivalent to the LOD score). On chromosome 1, we have a major peak around marker *D1Rat90*, and a minor peak around marker *D1Mit12* (Fig. 2A). On the other hand, when we focus on the co-existence of the O/O genotype at two marker loci on chromosomes 1 and 17, we have a peak only around the marker pair *D1Mit12* × *D17Mgh2* (the black spot in Fig. 2C). Thus, consideration of the synergetic effects between markers is indispensable for analysis of multiple factors.

As the example shows above, multiple markers are involved in the regulation of glucose tolerance in the OLETF rat [10]. Besides the interaction among markers at a particular time point, their correlated behavior in a time series is another interest in relation to the glucose tolerance. For example, it is possible that there exist marker groups each of which is related with regulation of the glucose level at different time points after oral administration. By using a series of measurements of glucose level at time 0, 30, 60, 90, and 120 min after oral administration, we evaluated the behavior of markers in a time series.

2 Method

2.1 Association Studies

Conjunctive Rules. In a given dataset, association study tries to extract latent rules, for instance; *"If marker A is homozygous and marker B is also homozygous, then the trait value is high."*

	Φ_i	$m_{1,i}$	$m_{2,i}$	$\cdots$	$m_{M,i}$
1	Φ_1	$m_{1,1}$	$m_{2,1}$	$\cdots$	$m_{M,1}$
2	Φ_2	$m_{1,2}$	$m_{2,2}$	$\cdots$	$m_{M,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	Φ_N	$m_{1,N}$	$m_{2,N}$	$\cdots$	$m_{M,N}$

Figure 3: Dataset.

The dataset consists of genotype information at marker loci and the quantitative trait values of interest in each individual. If we let $m_{j,i}$ denote the genotype at jth marker locus in the ith individual, and Φ_i denote the trait value in the ith individual, the total data can be summarized in a table as Fig. 3 (M and N are the numbers of markers and individuals, respectively). In this study, we call the judgment whether or not the jth marker locus has the genotype v in the ith individual a *primitive test* on the jth marker and denote it as $(m_{j,i} = v)$.

To investigate the relations between particular genotypes at multiple marker loci and the quantitative trait value, we use a conjunctive rule as $G_i : (m_{j_1,i} = v_1) \times \cdots \times (m_{j_k,i} = v_k)$, where k is a given constant number and G_i returns true if all the primitive tests hold, otherwise returns false. The reason why we employed a conjunctive rule is that it can determine the correlated effects of the marker loci on the quantitative trait value. The traditional LOD score, on the other hand, misses the correlated effects since it essentially focuses on only the one-to-one relationships between a marker and the trait value.

Data Division. According to whether or not each individual satisfies the rule G_i, we divide the set of individuals S into two classes S_0 and S_1, letting S_0 and S_1 respectively consist of the individuals that do not satisfy G_i and those that satisfy it (we call this operation *data division by* G_i). Here, if the rule can sort out a class S_1 which contains most of the individuals with high trait values, then the marker loci which constitute the rule G_i are considered to be related to the trait.

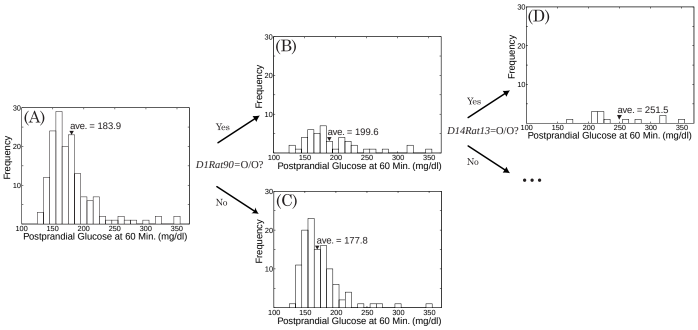

Figure 4: Data division in terms of genotype information.

Fig. 4 shows an example of the division of the population of 157 OLETF rats in terms of genotype information. Fig. 4A presents the original distribution of the blood glucose level in the population of 157 OLETF rats. Fig. 4B shows the distribution in the population of 44 rats which satisfy the rule (*D1Rat90* = O/O). The peak (130-220 mg/dl) in Fig. 4A is blunted and the average shifted to the right (183.9 to 199.6 mg/dl). Fig. 4C shows the distribution in the population of rats which do not satisfy the rule (*D1Rat90* = O/O). Fig. 4D shows the distribution in the population of 13 rats which satisfy the rule (*D1Rat90* = O/O) × (*D14Rat13* = O/O). The peak in Fig. 4A disappeared and the rats with poor glucose tolerance remained (average is 251.5 mg/dl).

2.2 Significance of a Rule

By using the distribution of the glucose level in the total population (S) and in two classes (S_0 and S_1), we can evaluate the significance of a rule. The F ratio defined as follows can be used for this purpose.

F Ratio. More generally, we consider that the total population S (including N individuals) is divided into k classes $(S_1, \ldots, S_k)$. Let μ and μ_i denote the average in S and S_i, respectively. *Mean square among classes* is defined as $MS_A = \frac{\sum_{i=1}^{k} |S_i|(\mu_i - \mu)^2}{k-1}$. *Mean square within classes* is defined as $MS_W = \frac{\sum_{i=1}^{k} \sum_{x \in S_i} (x-\mu_i)^2}{\sum_{i=1}^{k}(|S_i|-1)} = \frac{\sum_{i=1}^{k} \sum_{x \in S_i} (x-\mu_i)^2}{n-k}$. F ratio is calculated as the ratio of the mean square among classes (MS_A) to the mean square within classes (MS_W):

$$F = \frac{MS_A}{MS_W} = \frac{\sum_{i=1}^{k} |S_i|(\mu_i - \mu)^2/(k-1)}{\sum_{i=1}^{k} \sum_{x \in S_i} (x - \mu_i)^2/(n-k)}. \tag{1}$$

Intuitively, the F ratio takes a great value when the difference between the average in each class and total average is large (equivalently, when the variance in each class is small). F ratio greater than a statistical threshold constitutes evidence for a QTL. In this article, *maximization of the F ratio* means finding the conjunctive rule which maximizes the F ratio.

Theorem 1 *Maximization of the F ratio is equivalent to that of MS_A.*

Proof. We have $\sum_{x\in S}(x-\mu)^2 = \sum_{i=1}^{k}\sum_{x\in S_i}(x-\mu_i)^2 + \sum_{i=1}^{k}|S_i|(\mu_i-\mu)^2$, since $(x-\mu)^2 = \{(x-\mu_i)+(\mu_i-\mu)\}^2$. Therefore, $\sum_{x\in S}(x-\mu)^2 = (N-k)MS_W + (k-1)MS_A$. Thus, we also have:

$$F \quad = \quad \frac{MS_A}{MS_W} = \frac{(N-k)}{\frac{\sum_{x\in S}(x-\mu)^2}{MS_A} - (k-1)} \, \blacksquare \tag{2}$$

As Theorem 1 shows, the maximization of the F ratio is equivalent to that of MS_A. However, the F ratio is normalized using MS_W, therefore, the F ratio is more suitable for evaluation of the significance of data division with respect to the different phenotypes such as the glucose levels at time points in a time series. We can also prove that calculation of the F ratio of the data division in terms of the genotype at a single marker locus is essentially equivalent to that of the LOD score.

Theorem 2 (Nakaya et al. [10]) *Maximization of the LOD score is equivalent to that of MS_A.*

Proof. See Appendix ▮

Theorem 3 *Maximization of the F ratio is equivalent to that of the LOD score.*

Proof. Directly from Theorem 1 and Theorem 2 ▮
Theorem 3 gives an expansion of the definition of the LOD score which can evaluate effects of multiple markers, and also an interpretation of the F ratio in relation to the LOD score (see Appendix).

Based on the considerations as above, we tried to find the markers which maximize the F ratio with respect to each phenotype.

3 Search Program

Graph Search for Conjunctions. Consider all the conjunctions of the form $(m_{j_1,i} = v_1) \times \cdots \times (m_{j_k,i} = v_k)$, where $v_n = 0$ or 1. We first remark that it is NP-hard to compute the optimal conjunction that maximizes the F ratio [8]. One common approach to such optimization problems is an iterative improvement graph search algorithm that initially selects a candidate conjunction by using a greedy algorithm and then tries to improve the ensemble of candidate conjunctions by local search heuristics. To avoid the repetition of visiting the same node, conventional graph search algorithms maintain the list of visited nodes [1, 6], which however could be a severe bottleneck of parallel execution. We instead proposed a rule of rewriting a conjunction to others [9]. We first apply the rewriting rule to the initial conjunction to obtain child conjunctions, and then we repeat application of the rule to descendant conjunctions so that we can visit every conjunction just once without maintaining the list of visited conjunctions. Moreover, each application of the rewriting rule can be well parallelized.

For a dataset with a Boolean target attribute (e.g., indicating whether diseased or not), we have developed a branch-and-bound heuristics appropriate for the significance of correlation between a conjunction and the target attribute (expressed by χ^2 value) [9]. For a dataset with a numeric target attribute such as the glucose level, a similar heuristics based on the convexity of the mean square among classes (MS_A) is also available to prune the search space.

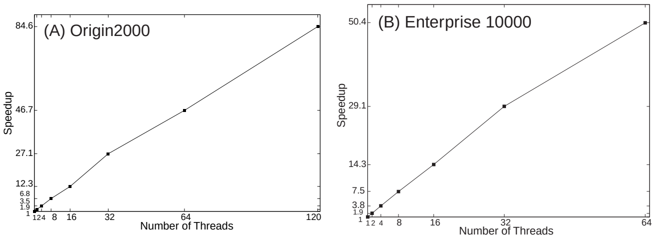

Figure 5: Speedup of calculation.

Implementation. We wrote a search program in the C++ language and parallelized it with the POSIX thread library on two commercially available parallel computers: the Sun Microsystems Enterprise 10000 (64 UltraSPARCII [250MHz] processors) and the SGI Origin2000 (128 R10000 [195MHz] processors). For a dataset of an intercross population, it can use primitive tests of the form ($m_{j,i}$ = O/O), ($m_{j,i}$ = F/F), ($m_{j,i}$ = O/F), and ($m_{j,i}$ = O/O or O/F) to reveal the dominant and the recessive effects of the marker loci.

In this work, we focused on the effects of the combinations of two markers. Therefore, we executed the program under the restriction $k = 2$. We calculated the F ratio of the data division by all the combinations of two markers without the branch-and-bound heuristics. During parallel execution, we used calculation of the F ratio of the data division by each marker combination as the unit of computing, since they can be carried out simultaneously. To distribute the computing among multiple P processors we statically divided the set of unit of computing into disjoint P subsets evenly and assigned them to the processors. Each processor iterates the calculation of the F ratio indicated in the assigned subset. When all the processors complete the calculation the program terminates. The required computation time for calculation of the F ratio for all the combinations of two markers using the dataset with 157 individuals and 279 markers is 16 seconds (Origin2000) and 37 seconds (Enterprise 10000). These results correspond to an 85-fold and a 50-fold calculation speedup, respectively. Calculation speedup scaled almost linearly with respect to the number of the processors used. Fig. 5 presents the relation between the number of the generated threads and the calculation speedup on the SGI Origin2000 and the Sun Microsystems Enterprise 10000.

4 Results

To find significant conjunctive rules with respect to oral glucose tolerance (measured as the postprandial blood glucose level after oral administration of glucose solution) we calculated F ratio of the data division by all the combinations of the k markers out of 279 markers ($k = 1$ and 2). As the target phenotypes, we used the blood glucose levels at time 0, 30, 60, 90, and 120 min after oral administration. We focus on rules which use the primitive tests of the form $(m_{j,i} = O/O)$. For simplicity, we denote a rule $G_i = (m_{j_1,i} = O/O) \times \cdots \times (m_{j_k,i} = O/O)$ as $m_{j_1} \times \cdots \times m_{j_k}$.

Single Marker. Fig. 6 shows the F ratio of the data division according to whether or not a single marker has the O/O genotype at each marker locus along chromosomes 1, 5, 7, 14, 16, 17, and X. In Fig. 6, one column corresponds to one chromosome and shows the curves of the F ratio along the chromosome with respect to the glucose levels at time 0, 30, 60, 90, and 120 min after oral administration from top to bottom.

In terms of glucose level at 60 min, for example, we can observe the peaks of the F ratio in the region around markers *D1Rat90*, *D7Wox6*, and *D14Mit5*. These three marker loci on chromosomes 1, 7, and 14 correspond respectively to the significant loci designated *Dmo1*, *Dmo2*, and *Dmo3* which have been found by traditional LOD score analysis [5]. We also have a peak on chromosome 14 near marker *Cckar* 34.9cM apart from *D14Mit5*. This marker also has been known to be related to the *Cc-kar* gene [5, 11]. The weak peak around *D1Mit12* on chromosome 1 and *DxMgh2* on chromosome X are also known to be linked to the oral glucose tolerance [5, 11].

The vertical direction in Fig. 6 shows the behavior of markers in a time series after oral administration. We can observe that all the markers do not have the peaks of the F ratio at the same time point after oral administration. For instance, *D1Rat90* on chromosome 1 has a peak ($F = 30.2$) at

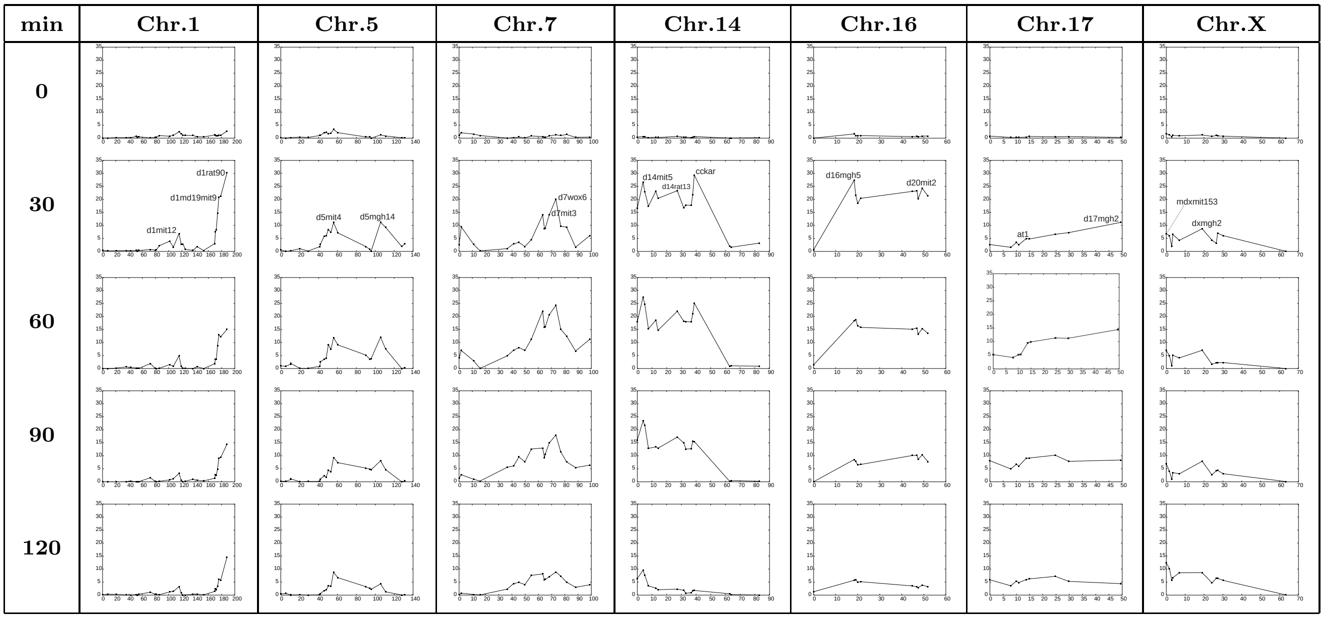

Figure 6: The F ratio of the data division in terms of the genotype at each marker locus along chromosomes (1, 5, 7, 14, 16, 17, and X) at 0, 30, 60, 90, and 120 min after oral administration. The horizontal axis indicates the genetic distance (cM) between markers and small rectangular marks show marker locus positions.

time 30 min, while *D7Wox6* on chromosome 7 has a peak ($F = 24.3$) at time 60 min, *MdxMit153* on chromosome X has a peak ($F = 12.4$) at time 120 min. *Cckar* on chromosome 14 has a peak at time 30 min, and keeps relatively high F ratios until 90 min. However, *Cckar* is not activated at 120 min. On the other hand, *D1Rat90* also has a peak at 30 min, but it is still activated at 120 min.

Two Markers. With respect to each phenotype (i.e., the glucose levels at time 0 to 120 min), we calculated the F ratio of the data division by all the possible conjunctive rules which consist of two markers, and then evaluated the effects of the co-existence of O/O genotypes at pairs of marker loci. According to the calculated F ratio we sorted the rules and picked up the pairs of chromosomes on which exist marker pairs with high F ratios. With respect to the glucose level at time 60 min, for example, we have significant pairs of markers on the pairs of chromosomes: 1×14, 7×14, 17×14, 1×17, 7×X, and 1×5. Table 1 lists a part of the pairs of markers with a high F ratio on those chromosome pairs. For glucose level at 30 min, we have significant pairs of markers on chromosomes 1×14. For glucose level at 90 min, we have significant pairs of markers on chromosomes 1×14, 1×17, 7×14, 7×X, 14×17, 14×X, and 17×X. For glucose level at 120 min, we have significant pairs of markers on chromosomes 1×14, 1×17, 7×X, 14×17, and 14×X.

Table 1: The pairs of markers significant to oral glucose tolerance at time 60 min.

Chr×Chr	$m_{j_1} \times m_{j_2}$	F	$\|S_1\|$	μ_1	$\|S_0\|$	μ_0
1×14	*D1Rat90*×*D14Rat13*	58.5	13	251.5	144	177.8
1×14	*D1Rat166*×*Cckar*	50.9	12	251.0	145	178.4
7×14	*D7Wox6*×*D14Mit5*	56.0	13	250.4	144	177.9
17×14	*At1*×*D14Mit5*	55.7	8	270	149	179.3
1×5	*D1Md19Mit9*×*D5Mgh14*	49.3	13	247.4	144	178.2
7×X	*D7Wox6*×*DxMgh2*	51.7	11	254.6	146	178.6
1×17	*D1Mit12*×*D17Mgh2*	56.0	7	276.4	150	179.6

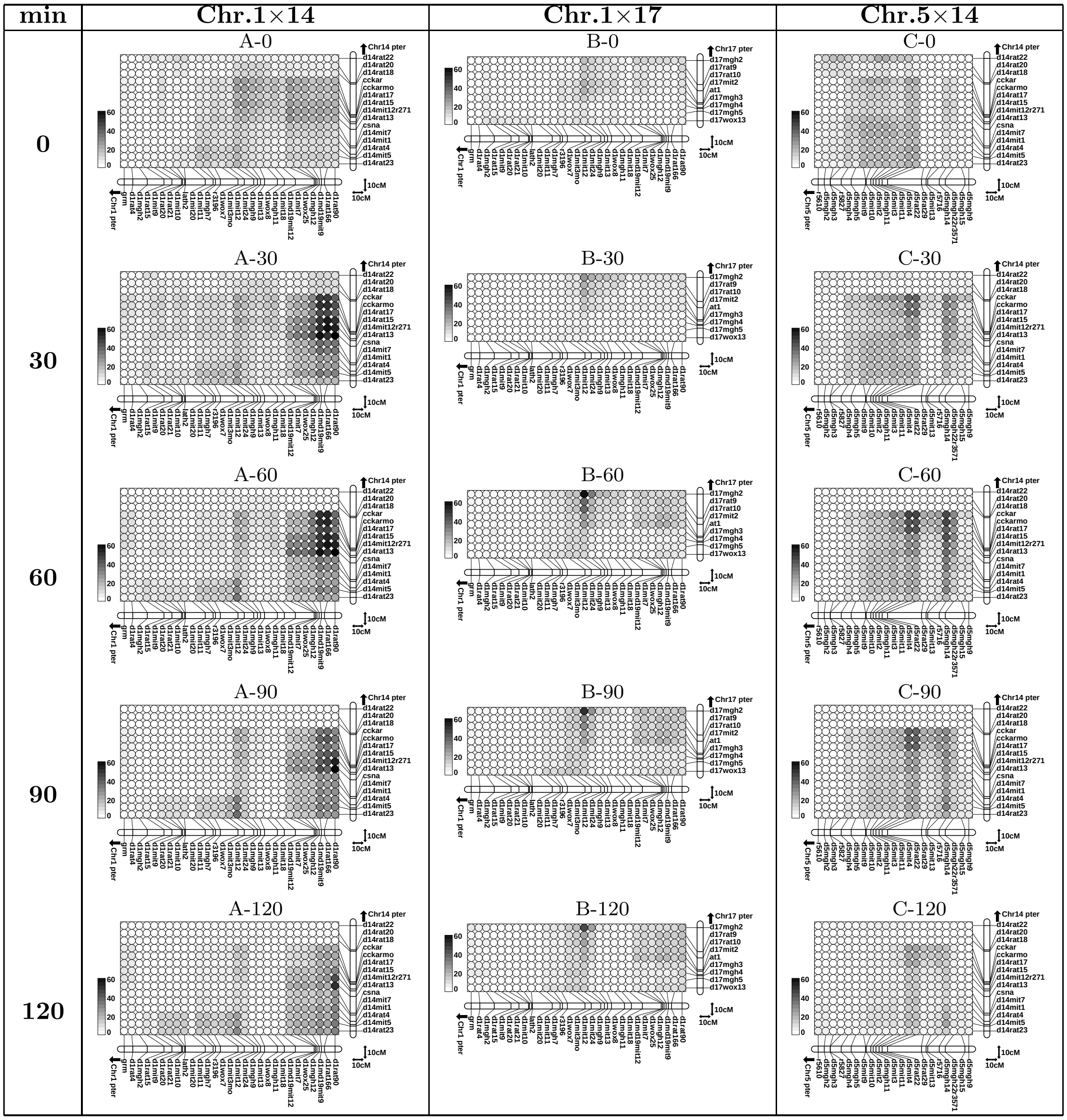

Figure 7: The F ratio of the data division by two markers.

In terms of the synergetic behavior of markers, at first glance, all the markers in the rules seem to have a peak of the F ratio even with a single marker alone (cf. Fig. 6), and each marker works in an additive manner. However, we can observe selectivity among the marker pairs. For example, the first four marker pairs in Table 1 use markers on chromosomes 1, 7, and 17 as the counterparts of those on chromosome 14. The three markers on chromosome 14 used in the pairs *D14Mit5*, *D14Rat13*, and *Cckar* exist in this order on chromosome 14, and their relative distances from *D14Mit5* are respectively 0, 23.3, and 34.9 cM. Other marker pairs on the chromosome pairs do not make the F ratio with respect to the glucose level at time 60 min high. This shows that a pair of markers does not make the F ratio high even if each of the two markers does alone.

Plotting the F ratio on a two-dimensional plane spanned by two chromosomes makes this clearer. Fig. 7 shows all the combinations of markers on the pairs of chromosomes (1×14, 1×17, 5×14, 7×14, 17×14, 7×X, and 16×14) at time 0, 30, 60, 90, and 120 min. A spot corresponds to a pair of markers and its color indicates the F ratio (dark one indicates a high F ratio). For instance, with respect to the

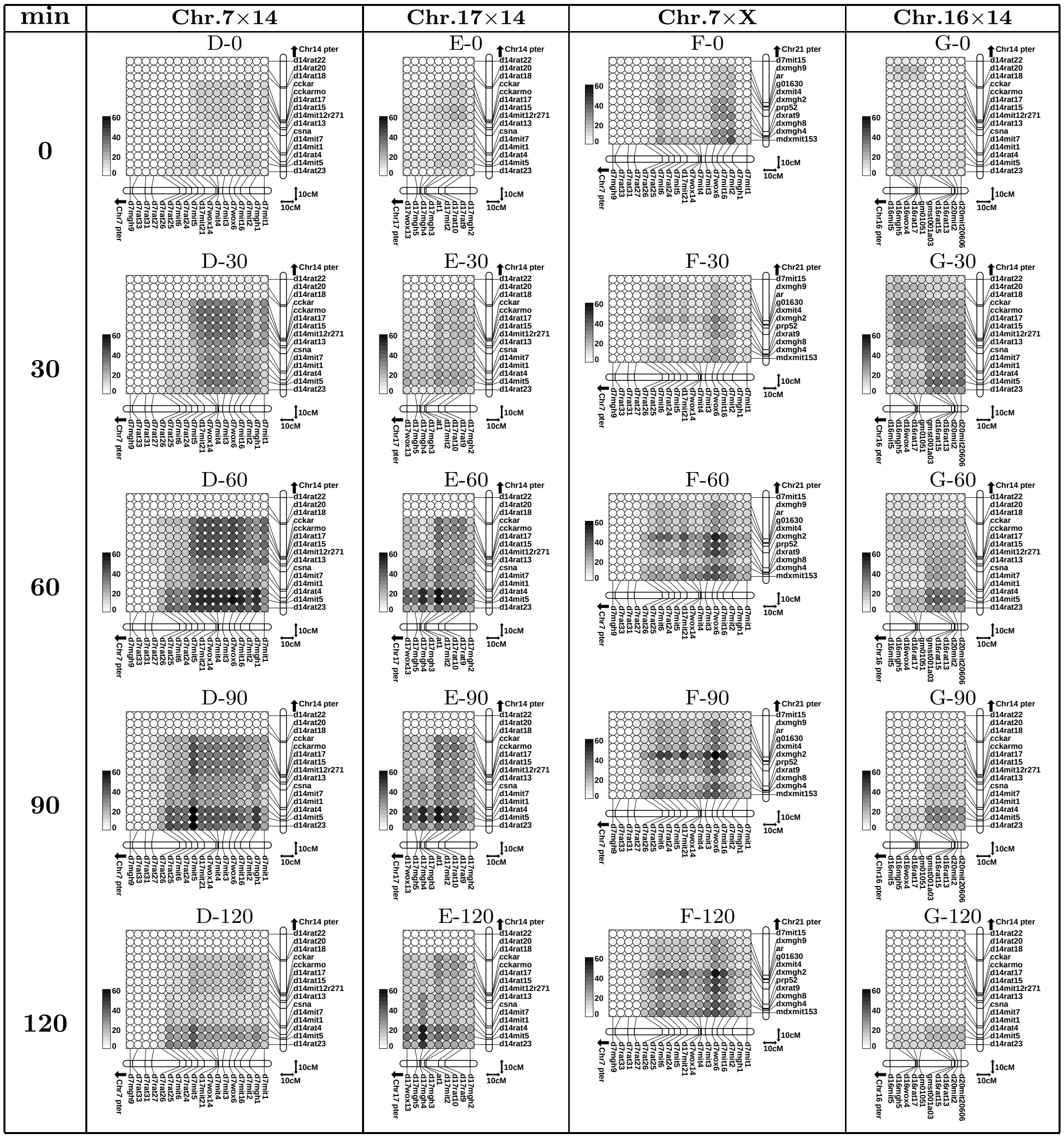

Figure 7 (cont.): The F ratio of the data division by two markers.

glucose level at 60 min, the markers in the region *D14Rat13–Cckar* on chromosome 14 make the F ratio high when they are paired with the those around *D1Rat90* on chromosome 1 (Fig. 7A-60). However, *D14Mit5* does not exert this influence under the same condition. On the other hand, *D14Mit5* makes the F ratio high when it is paired with the markers around *At1* on chromosome 17 but those in the region *D14Rat13–Cckar* do not (Fig. 7E-60). When the markers on chromosome 14 are paired with those on chromosome 7, pairs of markers work almost in an additive manner and we can observe two peaks in the region corresponding to the *D14Rat13–Cckar* and around *D14Mit5* (Fig. 7D-60). A similar situation can be found on chromosome 1. When the markers on chromosome 1 are paired with those on chromosome 17, the significant region around *D1Rat90* (Fig. 7A-60) disappears and another peak appears around *D1Mit12* × *D17Mgh2* instead (Fig. 7B-60).

All the marker pairs do not have the peaks of the F ratio at the same time point after oral administration. For example, on chromosomes 1×14, a significant region (i.e., in which the F ratio is high) appears from 30 min (Fig. 7A-30) while such a region appears from 60 min (Fig. 7B-60) on chromosomes 1×17. Even on a pair of chromosomes, the significant region changes in a time series.

For example, on chromosomes 17×14, the region *D17Mgh4*×(*D14Mit5*–*D14Rat4*) is still activated after the region *At1*×(*D14Mit5*–*D14Rat4*) is not activated (Fig. 7E-60,90,120). On the other hand, in relation to chromosomes 7×14, we can observe a siginificant region widely spead on the plane at time 60 min (Fig. 7D-60). However, a narrow siginificant region appears around *D7Mit5*×(*D14Rat23*–*D14Rat4*) at time 90 min (Fig. 7D-90).

5 Conclusion

We have focused on the relation between the multiple marker loci and the quantitative trait in a time series. To investigate the effects of multiple marker loci on the phenotype at each time point, we divided the set of the individuals into two classes according to a judgement whether or not each individual has particular genotypes at multiple marker loci. We formalized the judgement regarding genotypes as a conjunctive rule and estimated the significance of the rule in terms of the F ratio of its data division. The proposed method can determine the significant combinations of marker loci in relation to the phenotype at each time point by finding the rule accompanied by a high F ratio. We also showed that finding the significant marker loci based on the F ratio is equivalent to that based on the traditional LOD score.

The application of the above method on the OLETF model rat of non-insulin dependent diabetes mellitus (NIDDM) has found the combinations of marker loci significant to oral glucose tolerance (OGT) in a time series after oral administration of glucose solution. Plotting the F ratio on a two-dimensional plane spanned by two chromosomes presents clearly the relation among the marker loci. Observation of the F ratio on the plane with respect to measurements at each time point, can present the selectivity in the effects of the marker combinations in the time series. This property of the marker loci cannot be discovered solely by analysis of one-to-one relationships between a marker locus and the quantitative trait, as seen in the calculation of the LOD score along chromosomes. Thus, we have proposed a new method of QTLs analysis and showed its usefulness using experimental results in conjunction with real data.

Acknowledgments

This research is partly supported by a Grant-in-Aid for Scientific Research on Priority Areas "Genome Science" from the Ministry of Education, Science and Culture, Japan.

References

[1] Ananth, G.Y., Kumar, V., and Pardalos, P., *Parallel Processing of Discrete Optimization Problems*, John Wiley & Sons, 1993.

[2] Galli, J., Li, L.S., Glaser, A., Ostenson, C.G., Jiao, H., Fakhrai-Rad, H., Jacob, H.J., Lander, E.S., and Luthman, H., Genetic analysis of non-insulin dependent diabetes mellitus in the GK rat, *Nature Genet.*, 12:31-37, 1996.

[3] Gauguier, D., Froguel, P., Parent, V., Bernard, C., Bihoreau, M.T., Portha, B., James, M.R., Penicaud, L., Lathrop, M., and Ktorza, A., Chromosomal mapping of genetic loci associated with non-insulin dependent diabetes in the GK rat, *Nature Genet.*, (12):38–43, 1996.

[4] Hirashima, T., Kawano, K., Mori, S., and Natori, T., A diabetogenic gene, ODB2, identified on chromosome 14 of the OLETF rat and its synergistic action with ODB1, *Biochem. Biophys. Res. Commun.*, 224:420–425, 1996.

[5] Kanemoto, N., Hishigaki, H., Miyakita, A., Oga, K., Okuno, S., Tsuji, A., Takagi, T., Takahashi, E., Nakamura, Y., and Watanabe, T.K., Genetic dissection of "OLETF", a rat model for non-insulin-dependent diabetes mellitus, *Mamm. Genome*, 9:419–425, 1998.

[6] Kumar, V., Grama, A., and Karypis, G., *Introduction to Parallel Computing: Design and Analysis of Algorithms*, Benjamin Cummings, 1993.

[7] Lander E.S. and Botstein, D., Mapping mendelian factors underlying quantitative traits using RFLP linkage maps, *Genetics*, 121:185–199, 1989.

[8] Morishita, S., On classificaton and regression, In *Proc. of Discovery Science, DS'98, Lecture Notes in Artificial Inteligence 1532*, 40–57, 1998.

[9] Morishita, S. and Nakaya, A., Parallel branch-and-bound graph search for correlated association rules, In *Proc. of Workshop on Large-Scale Parallel KDD Systems in conj. with the 5th ACM SIGKDD*, 25–34, 1999.

[10] Nakaya, A., Hishigaki, H., and Morishita, S., Mining the quantitative trait loci associated with oral glucose tolerance in the OLETF rat, In *Proc. of Pacific Symp. on Biocomp.* , 2000 (to appear).

[11] Takiguchi, S., Takata, Y., Funakoshi, A., Miyasaka, K., Kataoka, K., Fujimura, Y., Goto, T., and Kono, A., Disrupted cholecystokinin type-A receptor (CCKAR) gene in OLETF rats, *Gene*, 197:169–175, 1997.

[12] Watanabe, T.K., Bihoreau, M.T., McCarthy, L.C., Kiguwa, S.L., Hishigaki, H., Tsuji, A., Browne, J., Yamasaki, Y., Mizoguchi-Miyakita, A., Oga, K., Ono, T., Okuno, S., Kanemoto, N., Takahashi, E., Tomita, K., Hayashi, H., Adachi, M., Webber, C., Davis, M., Kiel, S., Knights, C., Smith, A., Critcher, R., Miller, J., James, M.R., *et al.*, A radiation hybrid map of the rat genome containing 5,255 markers, *Nature Genet.*, 22:27–36, 1999.

[13] Zeng, Z.-B., Precision mapping of quantitative trait loci, *Genetics*, 136:1457–1468, 1994.

Appendix: F Ratio and LOD Score

Here, we briefly explain the relation between the F ratio and the LOD score. We introduce the definition of the LOD score and show that finding peaks of the LOD score along the genome is equivalent to finding those of the F ratio.

Definition of the Lod Score. Assume that we have the genotypes of M marker loci and observational data of the quantitative trait in a population of N individuals. For a given marker locus, let g_i be an indicator variable which shows the genotype in the ith individual, and when the marker genotype in the ith individual has a particular genotype, g_i takes 1, otherwise it takes 0, and let Φ_i be the observation of the quantitative trait value in the ith individual. Using these variables we construct a regression model as $\Phi_i = a + bg_i + \varepsilon$ (a and b are regression coefficiencies). In this model, b corresponds to the effects of the genotype at the marker locus. ε is a normal variable with mean 0 and variance v. Using the assumption that the error term ε follows the normal distribution, the probability that ε takes a value x is defined as $z(x, v) = \frac{1}{\sqrt{2\pi v}} \exp\left(\frac{-x^2}{2v}\right)$. Thus, the probability that the observation of the quantitative trait values would have occurred under this parameterized model (likelihood) is $L(a, b, v) = \prod_{i=1}^{n} z(\Phi_i - (a + bg_i), v)$. We determine the unknown parameters a, b, and v so that this likelihood is maximized and let the solutions (called maximum likelihood estimators) be $\hat{a}$, $\hat{b}$, and $\hat{v}$, respectively. The obtained likelihood $L(\hat{a}, \hat{b}, \hat{v})$ is compared to the likelihood that the marker locus has no effect on the quantitative trait (i.e., $b = 0$). Let $L(\hat{\mu}, 0, \hat{v}_0)$ denote the latter likelihood ($\hat{\mu}$ and $\hat{v}_0$ are the average and the variance of the quantitative trait in the N individuals, respectively). The *LOD score* is the logarithm of the ratio: $\mathrm{LOD} = \log_{10}\left(\frac{L(\hat{a},\hat{b},\hat{v})}{L(\hat{\mu},0,\hat{v}_0)}\right) = \alpha(\ln L(\hat{a}, \hat{b}, \hat{v}) - \ln L(\hat{\mu}, 0, \hat{v}_0))$, where $\alpha = 1/ln10 > 0$ is a constant number. Note that the term $\ln L(\hat{\mu}, 0, \hat{v}_0)$ is constant. A LOD score greater than a statistical threshold constitutes evidence for a QTL. In this article, *maximization of the LOD score* means finding the marker locus which maximizes the LOD score.

Relation between LOD Score and F Ratio. We can prove that calculation of the F ratio of the data division in terms of the genotype at a single marker locus is essentially equivalent to that of the LOD score. Let S_0 and S_1 be the populations of individuals whose g_i is 0 and 1, respectively ($S_0 = \{i|g_i = 0\}$ and $S_1 = \{i|g_i = 1\}$). Let μ_0 and μ_1 denote the averages of Φ_i in S_0 and S_1.

Lemma 1 *Maximization of the LOD score is equivalent to that of* $\ln L(\hat{a}, \hat{b}, \hat{v}) = -n \ln \sqrt{2\pi} + -\frac{n}{2} \ln \hat{v} - \frac{n}{2}$*, where* $\hat{v} = \frac{1}{n} \sum_{i=1}^{n} (\Phi_i - (\hat{a} + \hat{b} g_i))^2$, $\hat{a} = \sum_{i \in S_0} \Phi_i / |S_0| = \mu_0$, $\hat{b} = \sum_{i \in S_1} \Phi_i / |S_1| - \sum_{i \in S_0} \Phi_i / |S_0| = \mu_1 - \mu_0$.

Proof. Likelihood is given by $L(a, b, v) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{v}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2v} \sum_{i=1}^{n} (\Phi_i - (a + b g_i))^2\right)$. Log-likelihood is given by $\ln L(a, b, v) = -n \ln \sqrt{2\pi} - \frac{n}{2} \ln v - \frac{1}{2v} \sum_{i=1}^{n} (\Phi_i - (a + b g_i))^2$. By differentiating the log-likelihood with respect to v, a, and b, and setting the derivatives equal to zero, we have the maximum likelihood estimators of v, a, and b as follows: $\hat{v} = \frac{1}{n} \sum_{i=1}^{n} (\Phi_i - (\hat{a} + \hat{b} g_i))^2$, $\hat{a} = \sum_{i \in S_0} \Phi_i / |S_0| = \mu_0$, $\hat{b} = \sum_{i \in S_1} \Phi_i / |S_1| - \sum_{i \in S_0} \Phi_i / |S_0| = \mu_1 - \mu_0$. Thus we have $\ln L(\hat{a}, \hat{b}, \hat{v}) = -n \ln \sqrt{2\pi} - \frac{n}{2} \ln \hat{v} - \frac{n}{2}$ ▮

Lemma 2 *Maximization of the LOD score is equivalent to minimization of* $\hat{v}$.

Proof. $LOD = \alpha(\ln L(\hat{a}, \hat{b}, \hat{v}) - \ln L(\hat{\mu}, 0, \hat{v}_0))$ where $\alpha > 0$ is a constant number. Since $\hat{v} \geq 0$, $\frac{dLOD}{d\hat{v}} = -\frac{\alpha n}{2\hat{v}} \leq 0$. Therefore, maximization of the LOD score is equivalent to minimization of $\hat{v}$ ▮

Lemma 3 *Minimization of* $\hat{v}$ *is equivalent to maximization of the mean square among classes* MS_A.

Proof. Let S_0 and S_1 be the sets of individuals whose g_i is 0 and 1, respectively, and let μ_0 and μ_1 denote the averages of Φ_i in S_0 and S_1. Let S be $S_0 \cup S_1$. We can rewrite $\hat{v} = \frac{1}{n} \sum_{i \in S_0} (\Phi_i - \mu_0)^2 + \sum_{i \in S_1} (\Phi_i - \mu_1)^2 = \frac{1}{n} (\sum_{i \in S} \Phi_i^2 - (|S_0| \mu_0^2 + |S_1| \mu_1^2))$. Since $MS_A = -n\mu^2 + (|S_0| \mu_0^2 + |S_1| \mu_1^2)$, minimization of $\hat{v}$ is equivalent to maximization of MS_A. Thus, maximization of the LOD score is equivalent to that of MS_A ▮

Theorem 4 *Maximization of the LOD score is equivalent to that of the* MS_A.

Proof. From Lemma 2 and Lemma 3 ▮

Handling Multiple Markers. Based on the properties above, we can define the LOD score that can evaluate the effects of multiple markers. Calculation of the F ratio of the data division by a rule $G_i : (m_{j_1,i} = v_1) \times \cdots \times (m_{j_k,i} = v_k)$ is equivalent to regression of the data on the model: $\Phi_i = a + bG_i + \varepsilon$.

Except for G_i, definitions of the variables are the same as those in the case with a single marker genotype g_i. In this regression model, coefficiency b is the phenotypic effect of the co-existence of particular genotypes $(v_1, \cdots, v_k)$ at multiple marker loci $(m_{j_1,i}, \cdots, m_{j_k,i})$. Here, we can use the same definitions of the likelihood function and the LOD score as explained previously, since the definitions do not depend on the definition of g_i. The significant combination of marker loci associated with the quantitative trait value is found by selecting a set of marker loci which maximizes the F ratio of the data division according to their genotype information. This provides a multi-dimensional expansion of the LOD score.