社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2006－MPS－62　(38)
2006－BIO－7　(38)
2006／12／22

遺伝子多型、環境因子を用いた多因子疾患の発症予測のための新規解析手法の開発

冨田康之 [1]、横田充弘 [2]、本多裕之 [1]

[1]名古屋大学大学院　工学研究科　化学・生物工学専攻
[2]愛知学院大学　歯学部　ゲノム情報応用診断学

心筋梗塞をはじめとする多因子疾患の発症は遺伝因子のみではなく、生活習慣などの環境因子の影響を大きく受けている。さらに、同じ疾患であっても発症要因は、個人により異なる。本研究では、(1)遺伝因子の組み合わせ、ある環境因子に対して感受性な遺伝因子の組み合わせによる危険因子の候補を複数提案し、(2)発症未知の人に対して、発症有無と要因を個人レベルで推定するモデル（環境因子と遺伝因子の組み合わせ）を構築する情報処理手法を開発した。選択された危険因子の中には、コレステロール値のような改善可能な環境因子を持つものが多数含まれており、疾患の一次予防への貢献が期待される。

Classification Method for Predicting the Development of Myocardial Infarction by Using the Interaction between Genetic and Environmental Factors

Yasuyuki Tomita[1], Mitsuhiro Yokota[2] and Hiroyuki Honda[1]

[1]Department of Biotechnology, School of Engineering, Nagoya University
[2]Department of Genome Science, School of Dentistry, Aichi-Gakuin University

In the present study, to predict the development of myocardial infarction (MI) and classify the subjects into personally optimum development patterns, we have extracted risk factor candidates (RFCs) that comprised a state that is a derivative form of polymorphisms and environmental factors using a statistical test. We then selected the risk factors using a criterion for detecting personal group (CDPG), which is defined in the present study. By using CDPG, we could predict the development of MI in blinded subjects with sensitivity greater than 80%. It can be an effective and useful tool in preventive medicine and its use may provide a high quality of life and reduce medical costs.

1. Introduction

The interaction between genetic and environmental factors, including diet and lifestyle, contribute to cardiovascular diseases, cancers, and other major causes of mortality. Myocardial infarction (MI), a cardiovascular disease, is generally caused by the occlusion of a coronary artery and is often induced by the rupture of a plaque, which occurs due to atherosclerosis of the coronary arteries. MI is a multifactorial disease that is caused due to complex interactions between various genetic and environmental factors on a polygenic basis [1]. The involvement of several environmental factors in the development of MI has been suggested; these include obesity, smoking, hypertension, diabetes mellitus, hypercholesterolemia and hyperuricemia [1]. In addition, the genetic factors responsible for the susceptibility to MI are believed to differ among patients based on environmental factors and other susceptible genes, despite the fact that the same disease is being considered. Therefore, it is very important to propose models that are a combination of various genetic and environmental factors that are associated with multifactorial diseases such as MI for the prediction of disease development and associated causes on an individual basis. This concept is useful for determining the treatment protocol for a patient and for disease prevention.

Methods with a high accuracy for the detection of the interaction between genes and the environment or between the genes themselves and for the prediction of the development of multifactorial diseases have rarely been proposed. Detection of these interactions by using conventional parametric statistical methods is difficult. Attractive and convenient tools showing an adequate level of performance should be established. In addition, stepwise forward selection, which is one of the methods for selecting reasonable variables, appears to omit important interactions of a combination that are statistically significant. The interaction containing only the first selected variable is selected, and the other significant interactions appear to be omitted. On the other hand, conducting an exhaustive search of the combined interactions of genetic and environmental factors by stepwise backward elimination is either impossible or time consuming if the model that is constructed first includes too many input variables. Similarly, it is impossible to select statistically significant factors when the sample size is relatively small.

Therefore, in the present study, first, exhaustive combinations comprising up to 3 factors were analyzed, and the risk factor candidates (RFCs) were extracted using binomial and random permutation tests. Second, the minimum number of risk factors from RFCs was selected and the development of MI was predicted in order to correctly classify not only the modeling data but also the blinded data by the criterion for detecting personal group (CDPG), which is defined in the present study [2]. The CDPG, our proposed method, was compared with AdaBoost (proposed by Freund and Schapire (1997)) and majority voting, which is whereby the option with a simple majority of votes wins. This is the first report on automatic selection of susceptible gene-gene and gene-environmental factor interactions in multifactorial diseases such as MI by using polymorphisms and environmental factors. For conducting a comparison of the performance of the CDPG, the personal developmental patterns of blinded data were analyzed by employing models constructed by using thousands of subjects. Further, to investigate the flexibility of this analysis, a 10-fold cross-validation was performed in RFC and risk factor selection processes.

2. Subjects and Methods

2.1 Subjects and data of polymorphisms and environmental factors

In our previous study, 22 and 20 polymorphisms were selected in males and females, respectively, from 112 common polymorphisms [1]. Candidate genes including these polymorphisms have been characterized and potentially associated with coronary atherosclerosis or vasospasm, hypertension, diabetes mellitus, or hyperlipidemia. The study population comprised 4152 Japanese subjects; 2460 subjects (1776 males and 684 females) had MI and 1692 subjects (1082 males and 610 females) did not exhibit any symptoms of MI. In the present study, the subjects with MI are referred to as “cases” and those without any symptoms of MI are referred to as “controls.” Since sex-based differences in the association between genetic polymorphisms and the risk of MI might be attributable, at least in part, to the differences in the levels of estrogen or other hormones between males and females, these were particularly analyzed.

Six environmental factors, namely, smoking, body mass index (BMI), hypertension, diabetes mellitus, hypercholesterolemia, and hyperuricemia, were used as the conventional risk factors for coronary artery disease. Their data were converted into binary data using a clinical protocol. In the present study, the subjects who smoked and those with hypertension, diabetes mellitus, hypercholesterolemia, and hyperuricemia are referred to as “positive” data, while the others are referred to as “negative” data. The subjects with and without obesity were classified based on their BMI as “high” and “low,” respectively. Each of the 1692 control subjects (1082 males and 610 females) had at least one “positive” or “high” data.

The data was divided into 10 groups by randomizing and alternating the data. Nine groups were assigned as modeling data, and 1 group was assigned as blinded data. Each group was assessed once as blinded data (10-fold cross-validation). Modeling data was used for combination analysis of gene-gene or genetic-environmental factors and for the selection of RFCs and risk factors mentioned later to predict the development of disease in blinded data and their classification into personal optimum development patterns. The more detailed information on data and their processing are shown in our report [2].

2.2 Extraction of RFCs

A binomial test was used to extract RFCs that might be associated with the development of MI. The test was performed in various combinations of up to 3 factors: (1) 1 polymorphism, 1 environmental factor, and (2) a combination of 1 polymorphism and 1 environmental factor; a combination of 2 polymorphisms, and (3) a combination of 2 polymorphisms and 1 environmental factor, a combination of 3 polymorphisms by using modeling data, except the missing data, that is, the subjects who had lost at least 1 of the polymorphism and environmental factor data in the combination. Combinations among environmental factors were not considered. The reason for employing this analysis was that we particularly considered the genes susceptible to each environmental factor related to the development of MI and the classification of each development pattern. The cause and effect relationship in the combinations was evaluated against exhaustive combinations of less than 3 of the factors mentioned above.

The most important cause and effect relationship among the combinations was defined as the remarkable rule (Figure 1) in which the existing ratio between the case and control is mostly biased among all combinations. The rule represents one square matrix in Figure 1; thus, in dominant or recessive analysis, there are 4 and 8 rules in case of 2 and 3 SNP combinations, respectively. For example, in rule 1 of Figure 1, subjects with the genotype AA of SNP A, B allele of SNP B, and negative state of the environmental factor are considered to be one of the rules for using the 2 SNP and 1 environmental factor combination.

Rule table				Polymorphism A	
				AA	Aa + aa
Polymorphism B	BB + Bb	Environmental factor	negative	rule 1 $N_{case,1}/N_{control,1}$	rule 2 $N_{case,2}/N_{control,2}$
			positive	rule 3 $N_{case,3}/N_{control,3}$	rule 4 $N_{case,4}/N_{control,4}$
	bb		negative	rule 5 $N_{case,5}/N_{control,5}$	rule 6 $N_{case,6}/N_{control,6}$
			positive	rule 7 $N_{case,7}/N_{control,7}$	rule 8 $N_{case,8}/N_{control,8}$

Figure 1. The rule table using a combination between 2 polymorphisms and 1 environmental factor. $N_{case,l}$ and $N_{control,l}$ represent the number of case and control subjects, respectively, belonging to rule l.

We assessed only one rule by using the P value mentioned below. The biased degree of relationship was evaluated with the existing ratio by the binomial test using the binomial distribution as follows:

$$f(N_{case,l}) = \frac{n!}{N_{case,l}!(n-N_{case,l})!} p^{N_{case,l}} (1-p)^{n-N_{case,l}} \quad (1)$$

where n is the sum of the observed number for $N_{case,l}$ and $N_{control,l}$ existing in rule l. The probability p represents N_{case} / (N_{case} + $N_{control}$), where N_{case} and $N_{control}$ represent the total number of cases and controls analyzed in the combination. The null hypothesis ($N_{case,l}$ / N_{case} $\leq$ $N_{control,l}$ / $N_{control}$) is tested by computing the sum (P value) of all $f(N_{case,l})$ that are equal to or lesser than that for the observed value of $N_{case,l}$ (one-tailed test)

Since there are 3 genotype patterns in each genetic factor, i.e., homozygote of the major allele, heterozygote, and homozygote of the minor allele in the SNP, the number of rules in a combination of 2 SNPs is 9. However, in the present study, since the method of SNP analysis using dominant and recessive concepts appears to be practical for the application of various phenotypes (such as diseases), the heterozygote is combined with either of the homozygotes mentioned below. Based on this information, data in high dimensions that is constructed by combining 3 genotype patterns can be reduced to lower dimensions by constructing it with combinations of the dominant and recessive genotype patterns and important evidence on the biological aspects might be obtained. The procedure for extraction of RFCs has been divided into 2 steps and is outlined in cited reference 2. In step 1, the *P* values were calculated from exhaustive genotype combinations of the dominant and recessive genotypes, for example, 2^g dominant and recessive combinations and $2^g \times 2^g$ rules in a combination of g SNPs. Then, a combination of dominant and recessive genotypes among the 2^g combinations was determined as a preferable combination for the prediction of MI, in which the *P* value in one of the rules under the condition $N_{case,l} / N_{case} > N_{control,l} / N_{control}$ was the lowest among the $2^g \times 2^g$ *P* values. The dominant model is a comparison of the Aa plus aa genotypes with the AA genotype, while the recessive model is a comparison of the aa genotypes with the AA plus Aa genotypes.

In order to extract RFCs, the statistical significance of the rule in each combination was assigned to the *P* value. In step 2, this was done by modeling the null distribution that had the lowest *P* value in each combination by using the random permutation test. In the random permutation test, the signal of the subject was randomized, thereby ensuring that the number of subjects in the rule did not change. We then examined how well the rule of correctly labeled data in each combination explains the extent of risk compared with the rule of randomly labeled data. The significance of the rule is $P^{ran}(P_x)$ (equation 2), which is the percentage of random rules.

$$P^{ran}(P_x) = \frac{1}{T_1 \times T_2} \sum_{i=1}^{T_1} \sum_{j=1}^{T_2} \theta(P_x - P_{i,j}) \qquad (2)$$

$\theta(z) = 1$ if $z \geq 0$, and it is equal to 0 otherwise. $P_{i,j}$ is the lowest *P* value of the rule obtained by using the randomly labeled data calculated with the binomial test in one combination and the permutation test in the other. P_x is the *P* value of the rule that uses correctly labeled data calculated with the binomial test. In other words, $P^{ran}(P_x)$ is the *P* value of P_x in the null distribution, which is the lowest *P* value in each combination, and this value is calculated using the random permutation test. T_1 and T_2 are the number of permutations and the number of combinations, respectively. In the present study, T_1 is 1000. T_2 is ${}_{22}C_2$ = 231 in the combination of 2 polymorphisms in males because in the random permutation test, the combination of dominant and recessive genotypes was already determined using the correctly labeled data mentioned above. In the present study, RFCs were inferred at the $P^{ran}(P_x)$ level by using this distribution and was calculated to be less than 0.01 ($P^{ran}(P_x) < 0.01$) by using a random permutation test.

2.3 Selection of risk factors from RFCs for the prediction of development and causal factors of blinded data

This section describes our new criterion, the CDPG [2], which is used for selecting the minimum number of risk factors in order to classify the blinded data into personally optimum development patterns and predict the disease development in these patterns. We refer to the RFCs that are selected by CDPG and other classification methods as “risk factors.” The

$$I = \frac{N^{(m)}{}_{RFC,case}}{N_{case}} - \frac{N^{(m)}{}_{RFC,control}}{N_{control}} \qquad (3)$$

selection of the m^th risk factor is carried out in order to maximize the index *I*.

$N^{(m)}{}_{RFC,case}$ and $N^{(m)}{}_{RFC,control}$ represent the number of case and control subjects who have more than 1 RFC while selecting the mth risk factor. N_{case} and $N_{control}$ represent the number of case and control subjects, respectively, in the modeling data, which adjust the difference of the number of subjects between cases and controls. Accuracy (*Ac*), sensitivity (*Se*), and specificity (*Sp*) in the selected M risk factors are defined as follows:

$$Ac = \frac{N^{(M)}_{RFC,case} + (N_{control} - N^{(M)}_{RFC,control})}{N_{case} + N_{control}} \quad (4)$$

$$Se = \frac{N^{(M)}_{RFC,case}}{N_{case}} \quad (5)$$

$$Sp = \frac{N_{control} - N^{(M)}_{RFC,control}}{N_{control}} \quad (6)$$

$N^{(M)}_{RFC,case}$ and $N^{(M)}_{RFC,control}$ represent the number of case and control subjects who had more than 1 risk factor among M risk factors. If the subject is a case and has more than 1 risk factor among M risk factors, the prediction is considered true (true positive; TP) and if the case subject has no risk factors, the prediction is considered false (false negative; FN). If the subject is a control and has no risk factor among M risk factors, the prediction is considered true (true negative; TN) and if the control subject has more than 1 risk factor, the prediction is considered as false (false positive; FP). The concept of selecting risk factors by the CDPG is employed to enable the selection of RFCs that would include more case subjects and less control subjects, preferably in the modeling data. Information on obtaining the execute code, for example, data and documentation of the CDPG software, is available at the following URL.
http://www.nubio.nagoya-u.ac.jp/proc/english/indexe.htm

We then compared our proposed method —CDPG—with 2 other classification methods, namely, AdaBoost and majority voting. In multifactorial disease, there might be no conclusive and sole risk factor for elucidating the developmental mechanism. The reason for employing these methods was that AdaBoost and majority voting have the same strategy for selecting input variables as CDPG. The strategy is that these methods predict the development of the disease with a focus on case or control subjects who can not be still explained with selected risk factors by selecting another risk factor stepwise [2].

The basic concept of AdaBoost is to repeatedly apply a simple learning algorithm called the weak learner to different weightings of the same training set (modeling data in the present study). In its simplest form, AdaBoost is intended for binary prediction problems where the training set consists of pairs $(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$; x_i corresponds to the features of an example and $y_i \in \{-1,+1\}$ is the binary label to be predicted. A weighting of the training examples is an assignment of a real value w_i to each example (x_i, y_i). Given a learning algorithm that generates a set of weak learners $h_1, h_2, \ldots, h_T$, the AdaBoost algorithms construct a combined hypothesis f of the form.

$$f(x) = \sum_{t=1}^{T} \alpha_t \cdot h_t(x) \quad (7)$$

α_t is the weight of the weak learner h_t, and both weights and hypotheses are learned by the AdaBoost algorithm. The final prediction learned by AdaBoost is sign [f(x)], which is weighted by majority voting ($f(x) > 0$: prediction result is case; $f(x) < 0$: prediction result, control). Majority voting is whereby the option with a non-weighted majority of votes wins. The difference between them is that selected risk factors are weighted or not [2].

3. Results and Discussion

3.1 Subjects and data of polymorphisms and environmental factors

In the present study, we analyzed 22 and 20 polymorphisms in 16 candidate genes of males and females, respectively, and 6 environmental factors as conventional risk factors for coronary artery disease. In males, diabetes mellitus had the lowest *P* value ($P = 2.18 \times 10^{-18}$) as a single factor. This was used as the sole factor for discriminating between the cases and controls in one of the modeling data sets. The accuracy of prediction was 52.9%, and the sensitivity and specificity were 34.3% and 83.5%, respectively, when the number of case subjects and control subjects were compared in order to assess the discrimination performance. Thus, sensitive prediction of disease development in all subjects by using a single factor was impossible, even though it had a statistically significant *P* value.

Therefore, initially, we focused on the combination analysis of polymorphisms and environmental factors. In data set 1 of males, in the 1 polymorphism-1 environmental factor combination, there were 80 RFCs; this constituted approximately 15% of the 528 rules, whereas in the combination of 2 polymorphisms, there were 18 RFCs; this constituted approximately 2% of the 924 rules. This tendency was observed in all data sets and females. Therefore, as analyzed in the present study, it is suggested that the development of MI might be more sensitive to environmental factors combined with polymorphisms that are susceptible to these factors. In addition, it is suggested that several risk factors that are susceptible combinations for the development of MI may be selected by a combination analysis of polymorphisms and environmental factors. Thus, we found that it was very important to analyze the combinations of polymorphisms and environmental factors for elucidating the mechanism of MI. In the present study, analyses of up to 3 combinations were performed because greater the number of factors constituting the combination, lesser the number of the subjects belonging to the rule and longer is the time required for the calculation. Therefore, the RFCs were used later for analysis. On the contrary, it is considered that subjects having MI comprise several groups in which the risk factors differ on an individual basis. Further, we selected susceptible risk factors for MI from RFCs to predict the development of MI in the subjects in the personal group. A personal group is a virtual group of individuals. We considered that all MI subjects are characterized by a pattern on the basis of which they can be classified into personal groups. We defined the CDPG that enables the classification of each group, including a large number of case subjects and few control subjects by restricting the number of risk factors to a minimum.

3.2 Selection of risk factors from RFCs and classification of blinded data into personal optimum development patterns

Our proposed method—CDPG—was compared with AdaBoost and majority voting as described in the Methods using MI model as well as a simulation study. The shift of *Ac*, *Se*, and *Sp* defined in the Methods is shown in Figure 2. A total of 30 risk factors were selected from the RFCs. We decided the number of risk factors when the *Ac* in modeling data averaged in 10-fold cross-validation reached the maximum value in the CDPG, AdaBoost, and majority voting (Table 1). In the CDPG model, the accuracy and sensitivity with both modeling and blinded data were high in males and females (Figure 2 and Table 1). In particular, sensitivity was high, indicating that the diagnosis of case subjects by using this model was more accurate than that with AdaBoost and majority voting. However, the specificity of our method was low as compared with that of AdaBoost and majority voting, indicating that the percentage of control subjects with a minimum of 1 risk factor was at least 40%. By using AdaBoost and majority voting, *Ac*, *Se*, and *Sp* hardly changed with risk factor selection in males and females.

Table 1. Accuracy, sensitivity, and specificity averaged in 10-fold cross-validation using risk factors selected by CDPG, AdaBoost, and majority voting.

(a) Males

modeling	CDPG	AdaBoost	majority voting
risk factors	28	3	3
accuracy	0.678	0.567	0.554
sensitivity	0.747	0.490	0.551
specificity	0.566	0.693	0.558

blinded	CDPG	AdaBoost	majority voting
accuracy	0.619	0.554	0.540
sensitivity	0.709	0.477	0.546
specificity	0.473	0.680	0.530

(b) Females

modeling	CDPG	AdaBoost	majority voting
risk factors	24	1	1
accuracy	0.736	0.631	0.631
sensitivity	0.824	0.430	0.430
specificity	0.638	0.856	0.856

blinded	CDPG	AdaBoost	majority voting
accuracy	0.645	0.631	0.631
sensitivity	0.751	0.429	0.429
specificity	0.527	0.857	0.857

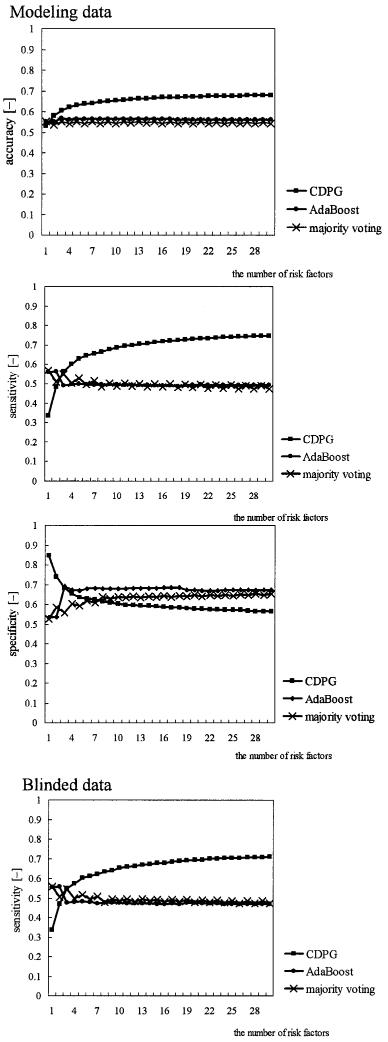

Figure 2. A shift in accuracy, sensitivity, and specificity in the procedure of selecting 30 risk factors with CDPG, AdaBoost, and majority voting in Males. Their values are averaged in 10-fold cross-validation.

3.3 Investigation of the extent of risk for each subject due to the interaction among risk factors

In CDPG analysis, by selecting a greater number of risk factors, the number of control subjects with a minimum of 1 risk factor and predicted to be case subjects increased (low specificity in CDPG). In case of multifactorial disease, the extent of risk for development appears to differ among the subjects. Although the specificity in CDPG was low, the extent of risk of control subjects might be lower than that of case subjects. Thus, in order to investigate the extent of risk for each subject, we paid attention to the interaction among the risk factors and examined it as follows.

By the CDPG method, 52.9% (572/1082) and 47.2% (288/610) of the male and female control subjects, respectively, of the blinded data have been assigned to the personal group through the 10-fold cross-validation by using the selected risk factors. Since it is believed that risk of development of a disease increases based on the interaction among the risk factors, we examined the relationship between the number of subjects and the number of risk factors (NRF) (Figure 3). The risk rate was defined as follows (equation 8).

$$Risk\ Rate = \frac{N_{case,\ NRF\ \geq R}}{N_{case,\ NRF\ \geq R} + N_{control,\ NRF\ \geq R}} \qquad (8)$$

R represents the cutoff value of NRF. $N_{case,NFR\geq R}$ and $N_{control,NFR\geq R}$ represent the number of case and control subjects who had more than R risk factors from the risk factors selected by the CDPG. The shift of risk rate in males is shown in Figure 3. It was observed that the risk rate was higher with increasing R in both of the modeling and blinded data. The same result was obtained in females, thereby satisfying the conditions R $\geq$ 3 and R $\geq$ 4 in males and females, respectively. The number of male and female subjects who had more than 4 and 5 risk factors, respectively, was less when compared with the total number of subjects in the modeling data (the number of case subjects was less than 20% of all the case subjects). When the cutoff value was defined as 3 and 4 in males and females, respectively, the respective risk rates were 76.1% and 76.8%. In the blinded data, the value was higher than the *Ac* (61.9% and 64.5% in males and females, respectively) which was defined as follows: if the subject has more than 1 risk factor among M selected risk factors, the prediction is case. Thus, it was observed that the interaction among risk factors selected by the CDPG had increased the risk of developing MI.

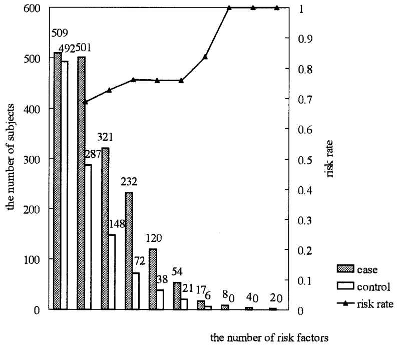

Figure 3. The number of risk factors that the case or control subjects have among 28 risk factors in males selected using CDPG and the number of subjects in 10 blinded data sets. Risk rate represents the rate of case subjects who have more than given number of risk factors.

In the present study, subjects with CC genotype at *TGF-βl* (T896C), CC or CT genotype at *ApoCIII* (C1100T) and positive of hypercholesterolemia were selected as one of the personal groups by CDPG and estimated to be at high risk for pathogenesis of MI in males (Figure 4). It has been assumed that subjects in this rule might be able to avoid development of MI by reducing cholesterol level.

(a)

				TGF β1 T869C (Leu10Pro)	
				TT + TC	CC
Apo CIII C1100T	CC + CT	Hypercholesterolemia	negative	383 /288	159 /101
			positive	296 /146	113 /38
	TT		negative	221 /167	70 /52
			positive	164 /98	60 /36

(b)

				TGF β1 T869C (Leu10Pro)	
				TT + TC	CC
Apo CIII C1100T	CC + CT	Hypercholesterolemia	negative	42 /44	19 /18
			positive	36 /13	14 /2
	TT		negative	24 /16	9 /4
			positive	17 /7	9 /2

Figure 4. Polymorphisms and environmental factor combination between *TGF-βl* (T896C) and *ApoCIII* (C1100T) may be associated with MI (gray rule) in males. (a) modeling data and (b) blinded data in one of the data sets.

In conclusion, we were able to classify the case and control subjects into personally optimum development patterns for multifactorial diseases such as MI with a high accuracy. For this, we used risk factor combinations that were selected by the binomial test and the random permutation test, which analyzes exhaustive combinations between polymorphisms and environmental factors, and CDPG, our proposed method, which is defined in the present study. Therefore, the CDPG method can be an effective and useful tool in preventive medicine and its use can provide high quality of life and reduce medical costs.

References

[1] Yamada, Y., *et al.*, Prediction of the risk of myocardial infarction from polymorphisms in candidate genes, *N. Engl. J. Med.*, 347:1916-1923 (2002).

[2] Tomita, Y. *et al.*, Classification method for predicting the development of myocardial infarction by using the interaction between genetic and environmental factors., *IPSJ Transactions on Bioinformatics*, *in press* (2006).