Original Paper

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

Yasuyuki Tomita,[†1] Hiroyuki Asano,[†2] Hideo Izawa,[†2]
Mitsuhiro Yokota,[†3,☆] Takeshi Kobayashi[†4]
and Hiroyuki Honda[†1]

Multifactorial diseases, such as lifestyle-related diseases, for example, cancer, diabetes mellitus, and myocardial infarction, are believed to be caused by the complex interactions between various environmental factors on a polygenic basis. In addition, it is believed that genetic risk factors for the same disease differ on an individual basis according to their susceptible environmental factors. 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 an accuracy greater than 75%. In addition, the risk percentage for MI was higher with an increase in the number of selected risk factors in the blinded data. Since sensitivity using the CDPG was high, 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 [1)]. 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)~3)]. Recent genetic linkage and association studies have already identified several candidate genes that may be responsible for predisposition to MI [3),4)]. A novel genetic susceptibility locus for MI has been identified on the chromosomal region 1p34-36 with the modified Haseman-Elston regression model (LOD = 11.68) [5)]. Thus, genetic factors may be necessary for the development of MI; however, this disease will not manifest in the absence of an environmental risk factor [6)]. The involvement of several environmental factors (conventional risk factors for coronary artery diseases) in the development of MI has been suggested; these include hypertension, diabetes mellitus, hypercholesterolemia, hyperuricemia, obesity, and smoking [3)]. In our previous paper, a significant correlation has been observed on a sex-specific basis between the genotypes of *connexin37*, *plasminogen-activator inhibitor type 1*, and *stromelysin-1* genes and the risk of MI by logistic regression analysis after adjusting for age; body mass index; and the prevalence of smoking, hypertension, diabetes mellitus, hypercholesterolemia, and hyperuricemia in a large Japanese cohort. However, 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

†1 Department of Biotechnology, School of Engineering, Nagoya University
†2 Department of Cardiology, Graduate School of Medicine, Nagoya University
†3 Department of Cardiovascular Genome Science, School of Medicine, Nagoya University
†4 School of Bioscience and Biotechnology, Chubu University
☆ Presently with Department of Genome Science, School of Dentistry, Aichi-Gakuin University

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.

In a previous study, we used single nucleotide polymorphism (SNP) data and an artificial neural network (ANN) for the prediction of childhood allergic asthma that might be strongly influenced by genetic factors [7]. The study comprised 344 subjects with data for 25 SNPs; these data were analyzed, and the ANN model predicted the diagnosis with an accuracy that was higher than 74%. The accuracy achieved by using ANNs and the SNP data was considerably high. In the case of multifactorial diseases such as MI, the developmental mechanism appears to differ among individuals because of the involvement of many risk factors. The number of risk factors and their combination patterns might differ across patients with multifactorial diseases. In addition, if several significant interactions are identified, it is very difficult to predict whether the unknown subject might develop the disease in the future.

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. 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

Data of polymorphisms and environmental factors were kindly provided by the Department of Cardiovascular Genome Science, Nagoya University, School of Medicine, Japan. Using public databases, including PubMed and Online Mendelian Inheritance, candidate genes that have been characterized and potentially associated with coronary atherosclerosis or vasospasm, hypertension, diabetes mellitus, or hyperlipidemia were selected. This selection was done on the basis of a comprehensive overview of vascular biology, platelet and leukocyte biology, coagulation and fibrinolysis cascades, and lipid and glucose metabolism and other metabolic factors [3]. In our previous study, 22 and 20 polymorphisms of these genes were selected in males and females, respectively, from 112 common polymorphisms [3]. Most of these were in the promoter regions, exons, or splice donor or splice acceptor sites in introns, and they might possibly cause changes in the function or level of expression of the encoded protein (**Table 1**). In Table 1, the dominant or recessive form in each polymorphism is indicated by the lower P value, implying that the number of case subjects is more biased; this is because in this study, we paid greater attention

Table 1 Genes, polymorphisms, and environmental factors examined in the present study.

Males

(i) dominant

gene	polymorphism	AA[b]				Aa + aa[b]			
		P value[a]	case[c]	control[c]	prediction[d]	*P* value	case	control	prediction
APOCIII[e]	C1100T	0.2089	288	160	1	0.3608	1,437	878	0
APOE	e4	0.4205	1,287	826	0	0.3509	359	221	1
	e2	0.3754	1,556	977	1	0.1031	90	70	0
	C4070T (Arg158Cys)	0.3954	1,623	980	1	0.1496	99	71	0
	T3932C (Cys112Arg)	0.4175	1,304	828	0	0.3457	365	222	1
CX37	C1019T (Pro319Ser)	0.0440	1,150	747	0	0.0062	588	295	1
NOS3	T-786C	0.3457	1,372	845	0	0.2175	366	207	1
GNB3	C825T (splice variant)	0.2376	429	274	0	0.3390	1,305	774	1
P22	C242T (His72Tyr)	0.0861	1,410	805	1	0.0042	338	255	0
PLA2G7	G994T (Val279Phe)	0.1281	1,192	754	0	0.0462	578	308	1
THBD	C2136T (Ala455Val)	0.2790	923	567	1	0.2550	710	468	0
THPO	A5713G	0.1435	359	238	0	0.2880	1,359	804	1
THBS4	G1186C (Ala387Pro)	0.2332	1,396	927	0	0.0336	241	127	1
TNFA	C-863A	0.1298	1,293	746	1	0.0306	435	304	0

(ii) recessive

gene	polymorphism	AA + Aa[b]				aa[b]			
		P value[a]	case[c]	control[c]	prediction[d]	*P* value	case	control	prediction
AGT	G-6A	0.0131	651	339	1	0.0498	1,090	713	0
APOCIII	C-482T	0.3091	1,312	791	1	0.1770	365	243	0
APOE	G-219T	0.0832	798	530	0	0.0914	912	521	1
CCR2	G190A (Val64Ile)	0.4507	1,581	960	0	0.3457	157	90	1
GP1A	A1648G (Lys505Glu)	0.2038	164	92	1	0.3942	1,494	944	0
IL10	T-819C	0.2785	1,417	933	0	0.0577	213	114	1
	A-592C	0.2591	1,496	930	0	0.0417	226	112	1
TGFB1	T869C (Leu10Pro)	0.1553	1,197	795	0	0.0452	459	255	1

environmental factor	*P* value	state	case	control	prediction	*P* value	state	case	control	prediction
BMI	0.4220	low	1,465	900	0	0.3332	high	311	182	1
Smoking	0.1447	negative	750	486	0	0.1775	positive	1,026	596	1
Hypertension	7.06E-05	negative	941	462	1	9.29E-05	positive	835	620	0
Diabetes mellitus	9.56E-09	negative	1,160	907	0	2.18E-18	positive	616	175	1
Hypercholesterolemia	0.0011	negative	1,020	721	0	0.0001	positive	756	361	1
Hyperuricemia	0.1042	negative	1,542	891	1	0.0013	positive	234	191	0

(a) *P* value calculated by using the binomial test.

(b) AA or Aa + aa represents a dominant model, AA + Aa or aa represents a recessive model. The (i) dominant or (ii) recessive type is determined in lower *P* values, implying that the number of case subjects is more biased.

(c) The number of case and control subjects in all data.

(d) 0: The prediction result is control; 1: The prediction result is case. Both predictions are made using the binominal test.

(e) The symbol of gene was referred without abbreviating in Appendix section.

Females

(i) dominant

gene	polymorphism	AA				Aa + aa			
		P value	case	control	prediction	*P* value	case	control	prediction
APOE	e2	0.4770	582	549	1	0.4256	54	53	0
	e4	0.3261	476	464	0	0.2117	160	138	1
	T3932C (Cys112Arg)	0.3611	486	465	0	0.2631	160	139	1
	C4070T (Arg158Cys)	0.4570	604	550	1	0.3614	54	53	0
NOS3	T-786C	0.2737	546	488	1	0.0975	106	117	0
ET1	G5665T (Lys198Asn)	0.4554	320	297	1	0.4560	326	308	0
SELE	A561C (Ser128Arg)	0.4307	580	562	1	0.2611	40	45	0
GP1BA	C1018T (Thr145Met)	0.3782	512	462	1	0.2821	142	140	0
IRS1	G3494A (Gly972Arg)	0.3277	616	578	0	0.0214	38	20	1
IL6	C-634G	0.0833	392	337	1	0.0489	242	267	0
PAI1	4G-668/5G	0.0460	240	266	0	0.0802	392	337	1
TNFA	G-238A	0.4413	626	585	0	0.2315	28	21	1

(ii) recessive

gene	polymorphism	AA + Aa				aa			
		P value	case	control	prediction	*P* value	case	control	prediction
APOCIII	C-482T	0.3466	478	477	0	0.2316	146	130	1
TAP	G1051A (Arg219Lys)	0.3689	470	450	0	0.2883	176	155	1
CD14	C-260T	0.2562	460	433	0	0.1553	206	167	1
CX37	C1019T (Pro319Ser)	0.4437	636	585	0	0.2306	26	19	1
FABP2	G2445A (Ala54Thr)	0.2997	572	531	0	0.0853	94	68	1
PON1	G584A (Gln192Arg)	0.2868	562	541	0	0.0622	86	62	1
MMP3	5A-1171/6A	0.0970	170	142	1	0.2248	456	464	0
TNFA	C-850T	0.3498	620	591	0	0.0226	30	15	1

environmental factor	*P* value	state	case	control	prediction	*P* value	state	case	control	prediction
BMI	0.1198	low	536	514	0	0.0074	high	148	96	1
Smoking	0.1294	negative	582	555	0	0.0012	positive	102	55	1
Hypertension	0.4666	negative	288	255	1	0.4716	positive	396	355	0
Diabetes mellitus	6.62E-10	negative	390	522	0	2.65E-19	positive	294	88	1
Hypercholesterolemia	0.0113	negative	310	331	0	0.0119	positive	374	279	1
Hyperuricemia	0.3229	negative	598	548	0	0.1004	positive	86	62	1

Table 2 The number of subjects, polymorphisms, and environmental factors.

Males		Females	
2,858		1,294	
case	control	case	control
1,776	1,082	684	610
polymorphisms	environmental factors	polymorphisms	environmental factors
22 (16 genes)	6	20 (16 genes)	6

Table 3 The number of subjects in the modeling and blinded data in the 10-fold cross-validation process.

data set	Males				Females			
	modeling data		blinded data		modeling data		blinded data	
	case	control	case	control	case	control	case	control
1	1,599	973	177	109	616	549	68	61
2	1,601	971	175	111	623	542	61	68
3	1,597	975	179	107	612	553	72	57
4	1,599	973	177	109	620	545	64	65
5	1,598	974	178	108	607	558	77	52
6	1,597	975	179	107	610	555	74	55
7	1,599	973	177	109	622	542	62	68
8	1,592	980	184	102	611	553	73	57
9	1,590	983	186	99	616	548	68	62
10	1,612	961	164	121	619	545	65	65

to the dominant or recessive analysis.

The study population comprised 4,152 Japanese subjects; of these, 2,460 subjects (1,776 males and 684 females) had MI and 1,692 subjects (1,082 males and 610 females) did not exhibit any symptoms of MI (**Table 2**). The study protocol was approved by the committees on the ethics of human research of Nagoya University Graduate School of Medicine and Gifu International Institute of Biotechnology, and written informed consent was obtained from each participant. The subjects were diagnosed by experienced doctors [3]. 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 [8].

Six environmental factors, namely, habitual cigarette smoking, obesity (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 [3]. 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 (Table 1). Each of the 1,692 control subjects (1,082 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). The number of cases and controls in each data set is shown in **Table 3**. 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 in the manuscript to predict the development of disease in blinded data and their classification into personal optimum development patterns.

2.2 Extraction of RFCs

A binomial test was used to extract RFCs that might be associated with the development of MI. The case/control ratio was calculated for 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 mod-

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

Fig. 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.
The combination of dominant and recessive genotypes was determined when 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 P values calculated with exhaustive combinations of modeling data using the dominant and recessive concepts—2^2 combinations in this case.

eling 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 (**Fig. 1**) in which the existing ratio between the case and control is mostly biased among all combinations. The rule represents one square matrix in Fig. 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 Fig. 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. To date, several analytical approaches have been proposed for gene-gene interactions, including combinatorial partitioning method (CPM) [9] and multifactor dimensionality reduction (MDR) [10)~12)]; however, these approaches assess all rules together using a certain evaluation value for the interaction. For example, in the MDR method, gene-gene interactions are assessed by testing the accuracy in the 10-fold cross validation, the cross-validation consistency, and the P value computed by comparing its (accuracy or consistency) value with the empirical distribution. However, since a risk factor is considered to be composed of certain alleles or genotypes (one rule) in a combination and could be missed when all rules are assessed together, 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 [13)]:

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

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) [14)].

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

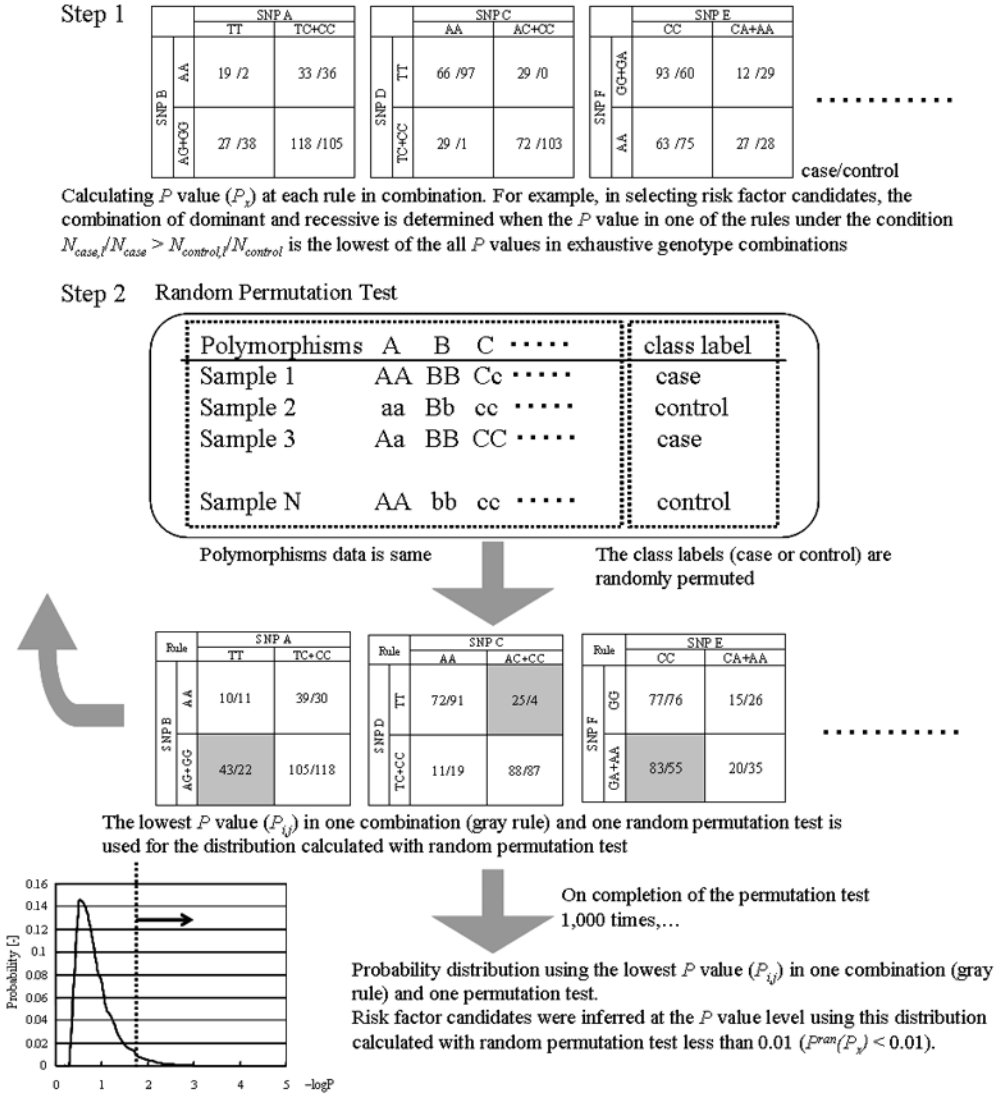

Fig. 2 The extraction procedure of risk factor candidates (RFCs) using the binomial test and the random permutation test in a combination with 2 polymorphisms.

been divided into 2 steps and is outlined in **Fig. 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. The P value of the polymorphisms and environmental factors analyzed in the present study is shown in Table 1.

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 [15)~17)]. 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 $P^{ran}(P_x)$ (Eq. (2)), which is the percentage of random rules [16].

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

$\theta(z) = 1$ 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 1,000. 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 Cover Rate and Case Rate

We defined the cover rate for each rule as the ratio of the subjects satisfying the rule to the total number of subjects and the case rate for each rule as the ratio of the case subjects to the subjects in the rule (Eqs. (3) and (4), respectively). $N_{case,l}$, $N_{control,l}$, N_{case}, $N_{control}$ are mentioned above.

$$CoverRate = \frac{N_{case,l} + N_{control,l}}{N_{case} + N_{control}}. \quad (3)$$

$$CaseRate = \frac{N_{case,l}}{N_{case,l} + N_{control,l}}. \quad (4)$$

Cover rate and case rate for the modeling data and blinded data, respectively, were calculated using a rule table that was constructed using modeling data.

2.4 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, 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 selection of the m^{th} risk factor is carried out in order to maximize the index I.

$$I = \frac{N^{(m)}_{RFC,case}}{N_{case}} - \frac{N^{(m)}_{RFC,control}}{N_{control}}. \quad (5)$$

$N^{(m)}_{RFC,case}$ and $N^{(m)}_{RFC,control}$ represent the number of case and control subjects who have more than 1 RFC while selecting the m^{th} 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:

$$Ac = \frac{N^{(M)}_{RFC,case} + N^{(M)}_{noRFC,control}}{N_{case} + N_{control}}. \quad (6)$$

$$Se = \frac{N^{(M)}_{RFC,case}}{N_{case}}. \quad (7)$$

$$Sp = \frac{N^{(M)}_{noRFC,control}}{N_{control}}. \quad (8)$$

$N^{(M)}_{noRFC,control} = N_{control} - N^{(M)}_{RFC,control}$. $N^{(M)}_{RFC,case}$ 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 mod-

eling 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 [18] 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.

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), ... , (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, ... , h_T, the AdaBoost algorithms construct a combined hypothesis f of the form,

$$f(x) = \sum_{t=1}^{T} \alpha_t \cdot h_t(x). \tag{9}$$

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, and *TP*, *FN*, *TN*, and *FP* are determined using $sign[f(x)]$ ($f(x) > 0$: prediction result is case; $f(x) < 0$: prediction result, control). In the present study, we constructed a weak learner h using RFC as follows. If the subject is a case and has the RFC rule in the combination among genes or genes and environmental factors, the prediction is considered true (*TP*); in other combinations, the prediction is considered false (*FN*). If the subject is a control and does not have the RFC rule in the combination, the prediction is considered true (*TN*); otherwise, the prediction is considered false (*FP*).

Majority voting is whereby the option with a non-weighted majority of votes wins. Its prediction result is as follows. If the subject is a case and its risk factor rate (RFR) is >0.5, the prediction is considered true (*TP*), otherwise, the prediction is considered false (*FN*). If the subject is a control and the RFR is <0.5, the prediction is considered true (*TN*), otherwise, the prediction is false (*FP*). RFR is m/M; m is the number of risk factors that the subject has among M selected risk factors. *Ac*, *Se*, and *Sp* are defined as follows:

$$Ac = \frac{N_{TP} + N_{TN}}{N_{case} + N_{control}}. \tag{10}$$

$$Se = \frac{N_{TP}}{N_{case}}. \tag{11}$$

$$Sp = \frac{N_{TN}}{N_{control}}. \tag{12}$$

N_{TP} and N_{TN} are the number of *TN* and *FN*, respectively. The criterion for selecting the risk factors by AdaBoost is to determine the hypothesis weight α_t, which is determined by minimizing the loss function in the modeling data [18], while that by majority voting is to determine the risk factors, which are determined by maximizing the *Ac* in the modeling data.

2.5 Data Simulation

To evaluate the power of CDPG for classifying subjects into personally optimum development patterns and predict the disease development in these patterns, we simulated case-control data including 10 development patterns and 1,000 non-development patterns. In the 10 patterns, the bias of case subjects with risk in all subjects had the same propensity as that of case subjects with selected RFCs derived from MI with respect to cover and case rates mentioned above ($0.1 < cover\ rate < 0.45$ and $0.7 < case\ rate < 1$). In the 1,000 non-development patterns, risk or not was randomly determined in case and control subjects; however, the cover and case rates of 1,000 patterns did not satisfy the propensity of the selected RFCs derived from MI. The simulation study population was the same as in the MI model in males comprising 2,858 subjects—1,776 case subjects and 1,082 control subjects. The case subjects had at least one risk in the 10 patterns. The number of control subjects without any development pattern was 293. The data was divided into 10 groups by randomizing and alternating

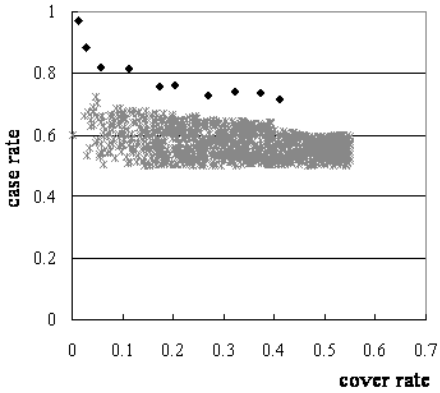

Fig. 3 Cover rate and case rate of simulation data. The 10 patterns (black dots) had the same propensity with respect to cover and case rates as those of the selected RFCs derived from MI (Fig. 4), while the 1,000 patterns (gray dots) did not have the same propensity as those of the selected RFCs.

the data. Nine groups were assigned as modeling data, and one group was assigned as blinded data. Each group was assessed once as blinded data (10-fold cross-validation). The number of cases and controls in each data set was the same as that shown in Table 3. Cover and case rates of the 1,010 simulation data are shown in **Fig. 3**. We also compared CDPG with 2 other classification methods, AdaBoost and majority voting. The criterion for selecting input variables by the 3 methods are shown above.

3. Results

3.1 Extraction of RFCs That Might be Associated with MI

Several reports have suggested an association between MI and genes or environmental factors; this association has been analyzed in the present study. For example, a novel genetic susceptibility locus for MI had been identified on chromosomal region 1p34-36 [5]. The C1019T polymorphism in the *connexin 37* gene lying on chromosomal region 1p34-36 was associated with a significant risk of MI in males, and the 4G-668/5G polymorphism in the *plasminogen-activator inhibitor type 1* gene and the 5A-1171/6A polymorphism in the *stromelysin-1* gene were associated with a significant risk of MI in females [3].

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 (Table 1). The study population and data sets for validation are shown in Tables 2 and 3. The association between these single factors and MI was assessed with $P^{ran}(P_x)$, which was calculated using the binomial test and the random permutation test described in the Methods. The number of extracted RFCs that were one of the rules (Fig. 1) in a combination that comprised genes and environmental factors with $P^{ran}(P_x)$ are shown in **Table 4**. For example, in data set 1, in males, there were 1 RFC and 44 rules when 1 polymorphism was used, while there were 3 RFCs and 12 rules when 1 environmental factor was used. The polymorphism is CT or TT of *connexin 37* (C1019T), and the environmental factors are negative for hypertension, positive for diabetes mellitus, and positive for hypercholesterolemia. In males, diabetes mellitus had the lowest P value as a single factor. This was used as the sole factor for discriminating between the cases and controls in modeling data set 1. 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. The procedure is outlined in Fig. 2. In data set 1, 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. 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. The same results were obtained in the case of females, as shown in Table 4 (ii). 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

Table 4 The number of risk factor candidates that satisfied the condition[a].

(i) Males

polymorphisms	environmental factors	all rules	risk factor candidates									
			data set	1	2	3	4	5	6	7	8	9	10
1	0	44		1	2	2	2	1	2	2	2	1	1
0	1	12		3	3	3	3	3	3	3	3	3	3
1	1	528		80	80	83	79	75	73	79	78	78	83
2	0	924		18	20	42	29	17	25	22	14	9	19
2	1	11,088		906	905	943	922	882	879	908	877	918	921
3	0	12,320		157	219	343	268	221	236	200	168	113	164

(ii) Females

polymorphisms	environmental factors	all rules	risk factor candidates									
			data set	1	2	3	4	5	6	7	8	9	10
1	0	40		0	0	1	0	1	2	0	1	0	0
0	1	12		2	4	3	4	3	4	4	3	3	4
1	1	480		57	79	57	74	70	67	56	58	62	65
2	0	760		5	9	12	8	24	29	8	16	5	14
2	1	9,120		594	790	649	700	718	709	546	622	643	673
3	0	9,120		103	150	150	117	221	228	84	169	121	143

(a) The P value is less than 0.01 ($P^{ran}(P_x) < 0.01$) when calculated by the binomial test and the random permutation test.

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 shown in Table 4 were used later for analysis.

In addition, RFCs were evaluated using 2 indices, namely, the cover rate and case rate. The former is the ratio of the subjects satisfying the rule to the total number of subjects and the latter is the ratio of the case subjects to the subjects in the rule (Eqs. (3) and (4) in the Methods). These ratios for males and females in data set 1 are shown in **Fig. 4**. RFCs had a characteristic propensity with respect to the cover rate and the case rate in the modeling data (Fig. 4). Since the case rates of all RFCs have been plotted against the cover rates of the RFC in Fig. 4, the correlation between the case rate and cover rate can be summarized. A risk factor that has a high value in both cover and case rates is indicative of disease development. However, when the development of MI was analyzed using polymorphisms and environmental factors, the observations were as follows: higher the cover rate, lower was the case rate (in 560 case subjects and 304 control subjects, the lowest case rate was 0.648 and the cover rate was 0.366), and lower the cover rate, higher was the case rate (in 13 case subjects and 0 control subjects, the lowest cover rate was 0.00541 and the case rate was 1 (Fig. 4 (a)). This tendency of RFCs was also observed in females (Fig. 4 (b)), and the blinded data also recorded similar results. Therefore, it was very difficult to select the most conclusive and sole risk factor for elucidating the developmental mechanism of MI.

On the contrary, it is considered that subjects

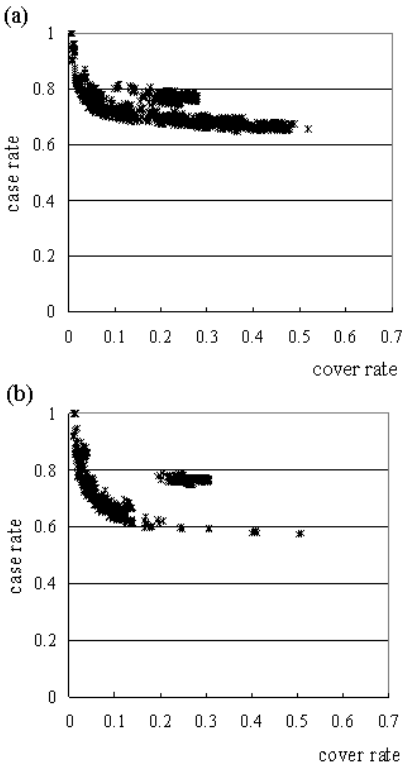

Fig. 4 Cover rate and case rate of risk factor candidates in modeling data set 1 of (a) Males and (b) Females.

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

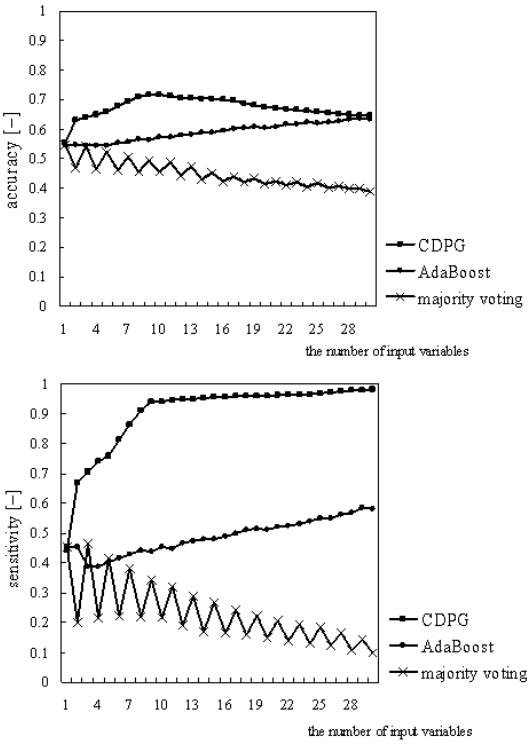

Fig. 5 A shift in the accuracy and sensitivity of blinded data in the procedure of selecting 30 input variables using CDPG, AdaBoost, and majority voting in simulation data. Their values are averaged in the 10-fold cross-validation.

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 Simulation Study

We performed a simulation study to evaluate the power of CDPG for classifying subjects into personally optimum development patterns and predicting the disease development based on these patterns. We investigated whether 10 development patterns can be selected from simulation data by means of CDPG. Simulation data are consisted of 1,000 variables that did not satisfy a propensity with respect to cover and case rates of the selected RFCs derived from MI model in males (Fig. 3). The shift of *Ac* and *Se* defined in the Methods in the case of blinded data is shown in **Fig. 5**. We counted the number of times which the 10 development patterns were selected within the first 10 selected input variables in each cross-validation step. The number was totaled in the 10-fold cross-validation and it was shown in **Table 5**. Accuracy, sensitivity, and specificity of blinded data using first selected 10 input variables are also shown in Table 5, which are averaged in the 10-fold cross-validation. As shown in Table 5, 10 development patterns were selected many times in CDPG as compared with that in AdaBoost and majority voting. One development pattern which had the highest cover rate was selected only once within the first 10 input variables in CDPG. On the other hand, in AdaBoost and majority voting, selected development patterns had a high cover rate. Although case subjects had at least one development pattern among the 10 patterns, the 2 methods could not classify subjects into each personal group.

Table 5 The number of times the 10 development patterns were selected within the first 10 selected input variables, and the accuracy, sensitivity, and specificity of blinded data using the first 10 selected input variables.

	CDPG	AdaBoost	majority voting
selected number (/100)	89[a]	39	36
accuracy	0.715[b]	0.572	0.456
sensitivity	0.941	0.453	0.319
specificity	0.345	0.765	0.850

(a) The number is totaled in the 10-fold cross-validation.
(b) The value is averaged in 10-fold cross-validation.

It was found that CDPG could classify subjects into personally optimum development patterns and predict the disease development in these patterns with high accuracy by selecting almost all development patterns. In addition, after the selection of the development patterns, the *Ac* in CDPG decreased with an increase in input variables; this is in contrast to that observed with AdaBoost. In conclusion, CDPG might be able to select various types of development patterns when there is no conclusive and sole risk factor for elucidating the developmental mechanism in multifactorial disease.

3.3 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 **Fig. 6**. A total of 30 risk factors were selected from the RFCs shown in Table 4 and those for males were different from those for females. We decided the number of risk fac-

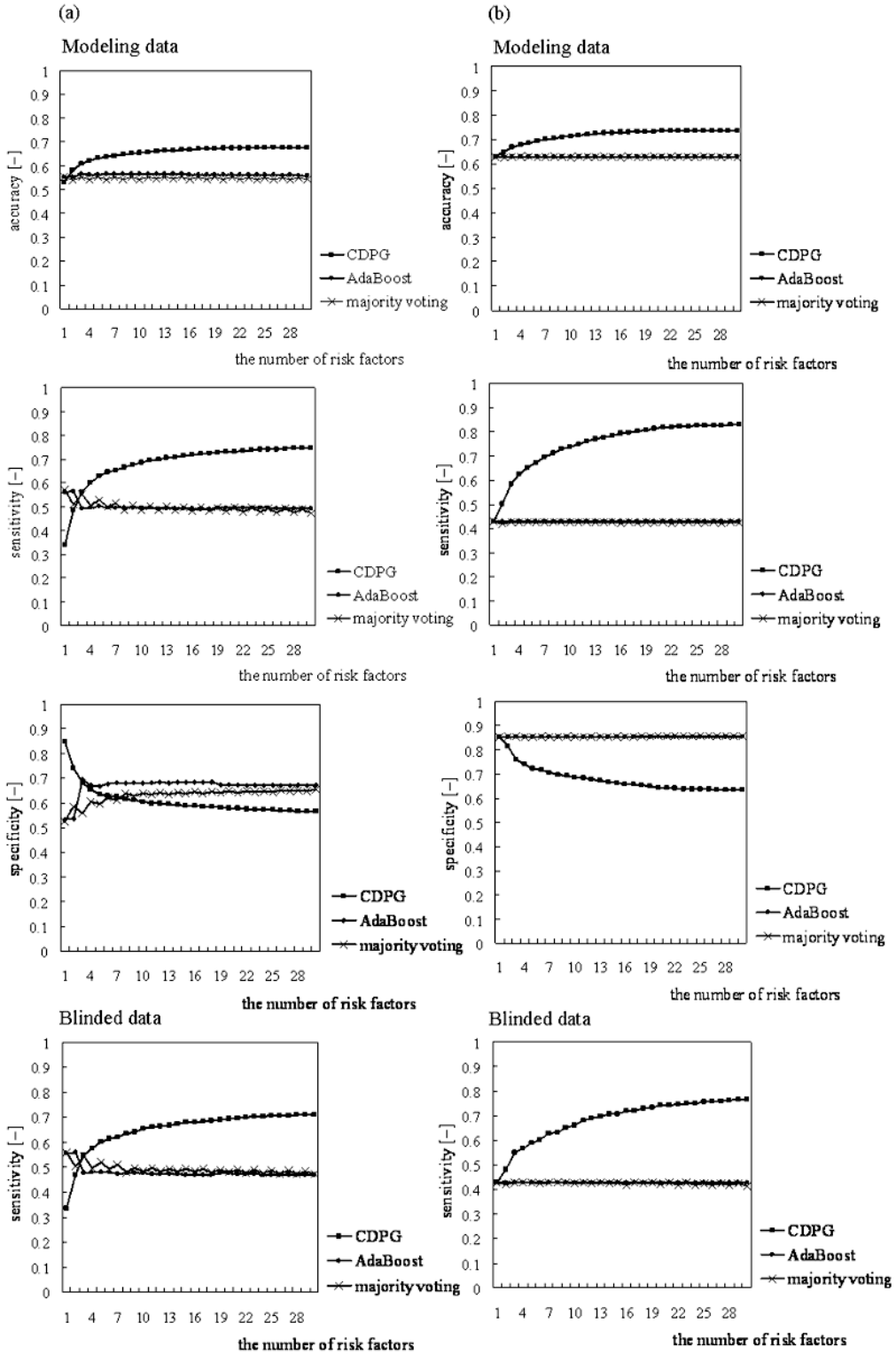

Fig. 6 A shift in accuracy, sensitivity, and specificity in the procedure of selecting 30 risk factors with CDPG, AdaBoost, and majority voting in (a) Males and (b) Females. Their values are averaged in 10-fold cross-validation.

Table 6 Accuracy, sensitivity, and specificity averaged in 10-fold cross-validation using risk factors selected by CDPG, AdaBoost, and majority voting.

(i) 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

(ii) 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

tors when the Ac in modeling data averaged in 10-fold cross-validation reached the maximum value in the CDPG, AdaBoost, and majority voting (**Table 6**). In the CDPG model, the accuracy and sensitivity with both modeling and blinded data were high in males and females (Fig. 6 and Table 6). In particular, sensitivity was high. When the risk factors were selected by the CDPG, the sensitivity of the prediction of case subjects in blinded data was 70.9% and 75.1% in males and females, respectively, whereas that of case subjects in the modeling data was 74.7% and 82.4% in males and females using 28 and 24 risk factors, respectively (Table 6), 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 (both males and females: approximately 60% and 50% in modeling and blinded data, respectively) 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.

When the risk factors were selected by AdaBoost, the sensitivity of the prediction of case subjects in blinded data was 47.7% and 42.9% in males and females, respectively, whereas that of case subjects in the modeling data was 49.0% and 43.0% in males and females, respectively. The number of risk factors selected by AdaBoost was 3 (mentioned below) and 1 (positive for diabetes mellitus) in males and females, respectively. When the risk factors were selected by majority voting, the sensitivity of the prediction of case subjects in blinded data was 54.6% and 42.9% in males and females, respectively, whereas that of case subjects in the modeling data was 55.1% and 43.0% in males and females, respectively. The number of risk factors selected by majority voting was 3 and 1 in males and females, respectively.

In data set 1, the 3 risk factors for males selected by AdaBoost were (1) negative for hypertension, (2) positive for diabetes mellitus and (3) a combination of CC of *p22phox* (C242T), AG or GG of *Thrombopoietin* (A5713G) and negative for hyperuricemia shown in **Table 7**. The cover rates of the 3 factors were 0.488, 0.276, and 0.497; their case rates were 0.676, 0.773, and 0.658. As shown in Table 7, the risk factors selected by the CDPG had higher case rates and lower cover rates when compared with those selected by AdaBoost. Since the CDPG is capable of classifying each group, which includes many case subjects and a few control subjects, by restricting the number of risk factors to a minimum (concept of the CDPG), it tends to select more risk factors that have low cover rates and high case rates compared with AdaBoost. This evidence of the power of our method for classification into each personal group was obtained in the simulation study. Thus, the CDPG achieves the concept mentioned above by selecting various risk factors that have high case rates and by decreasing the number of control subjects who have risk factors.

On the other hand, AdaBoost cannot select risk factors similar to those selected by the CDPG or those in the simulation study because weighting of the training w_i becomes larger while selecting these risk factors (i.e., the number of inaccurate classifications by using these weak learners (h_t) is higher). This concept is important when there is no conclusive risk factor that has high cover and case rate values. Therefore, a higher accuracy was achieved by using these risk factors compared with that obtained with factors selected by AdaBoost. The possibility that novel and significant factors for minor groups with respect to the development

Table 7 The number of male subjects who had a selected risk factor by (i) CDPG and (ii) AdaBoost in data set 1.

(i)

risk factor	modeling data		blinded data	
	case	control	case	control
1	528	148	66	13
2	282	104	34	20
3	162	61	25	7
4	281	107	33	19
5	263	104	38	20
6	47	12	2	3
7	24	1	4	0
8	113	38	14	2
9	18	2	0	0
10	42	8	5	1
11	31	6	1	1
12	29	4	1	0
13	73	21	9	4
14	43	11	5	4
15	45	11	3	2
16	26	4	1	4
17	32	7	2	2
18	17	1	2	0
19	30	6	7	3
20	32	6	3	0
21	19	2	0	1
22	33	7	4	0
23	76	19	10	1
24	110	39	12	3
25	20	2	3	0
26	341	90	42	10
27	14	0	1	0
28	76	14	9	2
more than one risk[a]	1,223	447	142	64
no risk[b]	376	526	35	45

(ii)

risk factor	modeling data		blinded data	
	case	control	case	control
1	848	407	93	55
2	548	161	68	14
3	842	437	100	53
more than one risk[a]	1,365	719	159	84
no risk[b]	234	254	18	25

(a) Subjects with a minimum of 1 risk factor and predicted to be case subjects.
(b) Subjects without any risk factor and predicted to be control subjects.

of multifactorial disease could be extracted using the CDPG was achieved in this study. Since minor risk factors (low cover rate but high case rate) might be present in multifactorial diseases, the CDPG is considered as an effective tool in terms of selecting risk factors when compared with AdaBoost and majority voting.

3.4 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% (572/1,082) 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) (**Fig. 7**). The risk rate (RR) was defined as follows (Eq. (13)).

$$RR = \frac{N_{case,NRF \geq R}}{N_{case,NRF \geq R} + N_{control,NRF \geq R}}. \tag{13}$$

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 is shown in Fig. 7. It was observed that the risk rate was higher with increasing R in the modeling data. The same result was obtained in the blinded data, 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.

By using CDPG, 18 and 16 risk factors were selected for (i) males and (ii) females, respectively, as shown in **Table 8**. These were selected at least 5 times by CDPG in the 10-fold

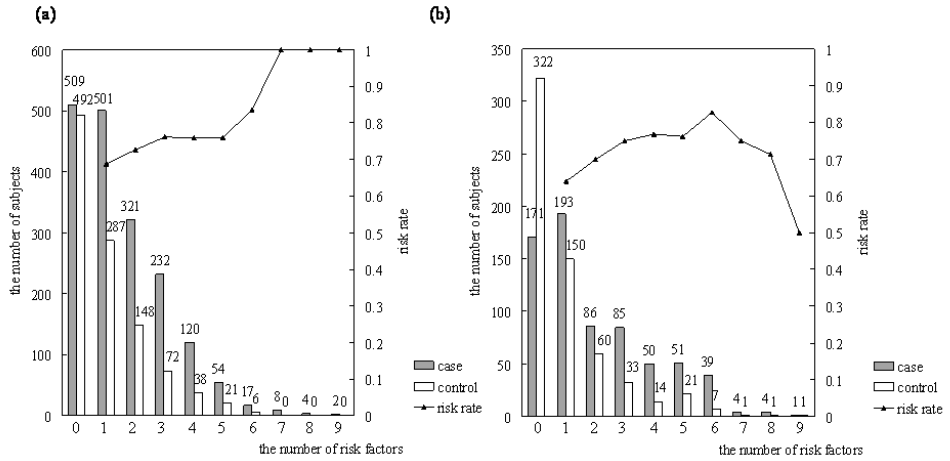

Fig. 7 The number of risk factors that the case or control subjects have among 28 (males) and 24 risk factors (females) selected using CDPG and the number of subjects in 10 blinded data sets of (a) Males and (b) Females. Risk rate represents the rate of case subjects who have more than given number of risk factors.

Table 8 Risk factors selected by CDPG.

(i) Males

gene polymorphism	genotype	*gene* polymorphism	genotype	*gene* polymorphism or environmental factor	genotype or state	n (/10)[a]
APOCIII C-482T	CC	*PLA2G7* G994T	GT + TT	BMI	high	10
APOCIII C1100T	CC	*AGT* G-6A	GG + GA	BMI	high	9
APOE e4	e3e4 + e4e4	*GP1A* A1648G	AA + AG	*AGT* G-6A	GG + GA	9
CX37 C1019T	CC + CT			Diabetes mellitus	positive	8
PLA2G7 G994T	TT			BMI	high	8
IL10 T-819C	TT	*IL10* A-592C	AC + CC			6
THBS4 G1186C	GC + CC	*THPO* A5713G	GG	Hypertension	negative	6
APOE e4	e3e3	*CCR2* G190A	AA	Hypertension	negative	6
THBD C2136T	CC	*APOCIII* C1100T	CC	*GNB3* C825T	TT	6
TGFB1 T869C	CC	*APOCIII* C1100T	CC + CT	*IL10* A-592C	CC	6
THBD C2136T	CC	*TGFB1* T869C	CC	Hyperuricemia	positive	5
THBS4 G1186C	GC + CC	*THPO* A5713G	GG	Smoking	positive	5
IL10 T-819C	TT	*IL10* A-592C	AC + CC	BMI	low	5
APOCIII C1100T	CC + CT	*GNB3* C825T	CT + TT	Hypercholesterolemia	positive	5
CX37 C1019T	CT + TT	*THPO* A5713G	AG + GG	*AGT* G-6A	GG + GA	5
IL10 T-819C	CC	*TGFB1* T869C	CC	*APOCIII* C1100T	CC + CT	5
GP1A A1648G	AA + AG	*APOCIII* C-482T	CC	*PLA2G7* G994T	GG	5
IL10 T-819C	CC	*TGFB1* T869C	TT	*APOE* G-219T	TT	5

(ii) Females

gene polymorphism	genotype	*gene* polymorphism	genotype	*gene* polymorphism or environmental factor	genotype or state	n (/10)[a]
				Diabetes mellitus	positive	10
MMP3 5A-1171/6A	5A5A + 5A6A	*TAP* G1051A	GG	Hypertension	negative	10
TNFA C-850T	CT + TT	*FABP2* G2445A	AA	*CD14* C-260T	TT	9
TAP G1051A	AA	*CD14* C-260T	TT	*ET1* G5665T	GT + TT	9
SELE A561C	AA + AC			Diabetes mellitus	positive	8
APOCIII C-482T	TT	*IL6* C-634G	CC	*GP1BA* C1018T	CC	8
CX37 C1019T	CC	*ET1* G5665T	GT + TT	BMI	high	7
PAI1 4G-668/5G	5G5G	*GP1BA* C1018T	CT + TT	Hypertension	positive	7
CX37 C1019T	CC	*FABP2* G2445A	AA	BMI	high	7
MMP3 5A-1171/6A	5A6A + 6A6A			Diabetes mellitus	positive	6
TAP G1051A	GG	*APOE* e4	e3e4 + e4e4	BMI	high	6
PON1 G584A	AA	*ET1* G5665T	GT + TT	Hypertension	negative	6
FABP2 G2445A	AA	*TAP* G1051A	AA	Hypertension	positive	6
MMP3 5A-1171/6A	5A5A + 5A6A	*PAI1* 4G-668/5G	4G4G	*APOE* e2	e3e2 + e2e2	6
PAI1 4G-668/5G	4G5G + 5G5G	*APOE* e2	e3e3	Hyperuricemia	positive	5
IL6 C-634G	CC	*FABP2* G2445A	GG	*TAP* G1051A	AA	5

(a) n represents the number of times the same combination comprised the genotype or the environmental factor was selected by CDPG in 10-fold cross-validation. Risk factors in which n was more than 5 are shown in this table.

cross-validation, and they showed the same patterns in terms of both their risk rule and dominant or recessive pattern.

4. Discussion

In the present study, initially, the analysis of exhaustive combinations of up to 3 factors was performed, and RFCs that had a statistically significant bias with regard to the number of case and control subjects and those that might be associated with MI were extracted using the binomial test and the random permu-

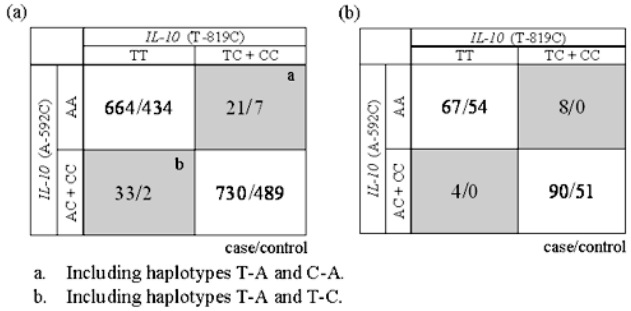

Fig. 8 Polymorphism combination between *IL-10* T-819C and A-592C may be associated with MI (gray rule) in males. (a) modeling data and (b) blinded data in data set 1.

tation test. Next, by analyzing the simulation data including 10 development patterns satisfying a propensity of RFCs, we obtained evidence of the high power of CDPG for classifying subjects into personally optimum development patterns and predicting the disease development in these patterns. As evidenced in the simulation study and MI model, we were able to classify the case and control subjects into personally optimum development patterns with a high accuracy.

Another objective of this step was to select the most susceptible RFCs and exclude the others that had the same pattern of development as mentioned above. For example, RFCs with cover and case rates that were more than 0.2 and 0.7 (Fig. 4), respectively, represent the RFCs, including one that was negative for hypertension, one that was positive for diabetes mellitus, and one that was positive for hypercholesterolemia in males and one that was positive for diabetes mellitus in females. As shown in Table 8, the CDPG selected the most susceptible risk factors, including the environmental risk factors, i.e., the most effective interaction between the environmental factors and genes in RFCs. In addition, since the risk rate for MI increased with an increase in the number of risk factors in both the modeling and blinded data, it was observed that the interaction among the risk factors selected by the CDPG had increased the risk of development of MI.

In preventive medicine, accurate prediction of subjects who might develop the disease(s) in the future and the development pattern of the disease is very important. In addition, warning these susceptible subjects regarding their risk factors is also necessary. Since the CDPG method showed a high sensitivity, it is considered as an effective and useful tool in preventive medicine and its use may provide a high quality of life and reduce medical costs.

To characterize the developmental mechanisms that are believed to differ among patients with multifactorial diseases such as MI based on their environmental factors and susceptible genes, despite the fact that the same disease was being considered, we investigated several relationships between polymorphisms and environmental factors that might not be exclusively associated with MI. In addition, based on their developmental mechanism, we classified the subjects into personal groups that comprised people who might have different susceptible factors related to MI.

In the present study, 2 risk factors comprised the TT genotype of *Interleukin-10* (*IL-10*) (T-819C) and AC or CC genotype of *IL-10* (A-592C), and the genotype combination with low for BMI were selected 6 and 5 times, respectively in 10-fold cross-validation process as one of the personal groups by CDPG and they were estimated to be at a high risk for the pathogenesis of MI (Table 8 (i), **Fig. 8**). Since the results of the modeling data were the same as those obtained in the blinded data set 1 (Fig. 8), polymorphisms in the promoter region of *IL-10* were found to be susceptible for MI and a responsible marker for MI in males.

In addition, the risk factor comprised the GT or TT genotype of *PAF-acetylhydrolase* (*PAF-AH*) (G994T), CC genotype of *Apolipoprotein CIII* (*ApoCIII*) (C-482T), and high for BMI were selected 10 times by CDPG in 10-fold cross-validation process as one of the personal groups, and they were considered to be at a high risk for the pathogenesis of MI (Table 8 (i), **Fig. 9**). It was reported that the T allele of *PAF-AH* G994T (Val279Phe) might exacerbate cardiac damage in Japanese individuals with

(a)

PAF-AH G994T		BMI	*ApoCIII* C-482T CC	*ApoCIII* C-482T CT+TT
GG		low	240 / 152	621 / 391
		high	48 / 35	122 / 76
GT+TT		low	106 / 70	283 / 159
		high	29 / 4	60 / 42

case/control

(b)

PAF-AH G994T		BMI	*ApoCIII* C-482T CC	*ApoCIII* C-482T CT+TT
GG		low	25 / 16	65 / 43
		high	6 / 4	17 / 10
GT+TT		low	11 / 10	35 / 18
		high	1 / 0	8 / 4

case/control

Fig. 9 Polymorphism and environmental factor combination among *PAF-AH* G994T, *ApoCIII* C1100T, and BMI might be associated with MI (gray rule) in males. (a) modeling data and (b) blinded data in data set 1.

hypertrophic cardiomyopathy [19]. The interaction between *PAF-AH*, *ApoCIII* and BMI for the development of MI is indicated in the present study.

As shown in Table 8, the risk factors that are combinations of polymorphisms and those that are susceptible to environmental factors were selected in both males and females. Using these risk factors, we were able to predict the development of multifactorial diseases such as MI and classify the subjects into personal optimum development patterns with a high accuracy by using candidate genes with known functions. Thus, it was very difficult to select the most conclusive and sole risk factor for elucidating the developmental mechanism of multifactorial disease such as MI. In conclusion, our proposed method–CDPG–that includes genetic and environmental factors can be an effective and useful tool because it enables the selection of various types of development patterns for MI and predicting the disease development in these patterns with high accuracy.

5. Conclusions

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.

Acknowledgments This work was supported in part by the Grant-in-Aid Scientific Research from the Ministry of Education, Culture Sports, Science and Technology of Japan (16012223, 17019028, 17209021 to Dr. Yokota). We also acknowledge THE HORI INFORMATION SCIENCE FOUNDATION for financial support.

References

1) Willett, W.C.: Balancing life-style and genomics research for disease prevention, *Science*, Vol.296, pp.695–698 (2002).
2) Marenberg, M.E., Risch, N., Berkman, L.F., Floderus, B. and de Faire, U.: Genetic susceptibility to death from coronary heart disease in a study of twins, *N. Engl. J. Med.*, Vol.330, pp.1041–1046 (1994).
3) Yamada, Y., Izawa, H., Ichihara, S., Takatsu, F., Ishihara, H., Hirayama, H., Sone, T., Tanaka, M. and Yokota, M.: Prediction of the risk of myocardial infarction from polymorphisms in candidate genes, *N. Engl. J. Med.*, Vol.347, pp.1916–1923 (2002).
4) Keavney, B., McKenzie, C., Parish, S., Palmer, A., Clark, S., Youngman, L., Delepine, M., Lathrop, M., Peto, R. and Collins, R.: Large-scale test of hypothesised associations between the angiotensin-converting-enzyme insertion/deletion polymorphism and myocardial infarction in about 5000 cases and 6000 controls, International Studies of Infarct Survival (ISIS) Collaborators, *Lancet*, Vol.355, pp.434–442 (2000).
5) Wang, Q., Rao, S., Shen, G.Q., Li, L., Moliterno, D.J., Newby, L.K., Rogers, W.J., Cannata, R., Zirzow, E., Elston, R.C. and Topol, E.J.: Premature myocardial infarction

novel susceptibility locus on chromosome 1P34-36 identified by genomewide linkage analysis, *Am. J. Hum. Genet.*, Vol.74, pp.262–271 (2004).
6) Broeckel, U., Hengstenberg, C., Mayer, B., Holmer, S., Martin, L.J., Comuzzie, A.G., Blangero, J., Nurnberg, P., Reis, A., Riegger, G.A., Jacob, H.J. and Schunkert, H.: A comprehensive linkage analysis for myocardial infarction and its related risk factors, *Nat. Genet.*, Vol.30, pp.210–214 (2002).
7) Tomita, Y., Tomida, S., Hasegawa, Y., Suzuki, Y., Shirakawa, T., Kobayashi, T. and Honda, H.: Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma, *BMC Bioinformatics*, Vol.5, No.120 (2004).
8) Guetta, V. and Cannon, R.O.: Cardiovascular effects of estrogen and lipid-lowering therapies in postmenopausal women, *Circulation*, Vol.93, pp.1928–1937 (1996).
9) Nelson, M.R., Kardia, S.L., Ferrell, R.E. and Sing, C.F.: A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation, *Genome Res.*, Vol.11, pp.458–470 (2001).
10) Hahn, L.W., Ritchie, M.D. and Moore, J.H.: Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions, *Bioinformatics*, Vol.19, pp.376–382 (2003).
11) Qin, S., Zhao, X., Pan, Y., Liu, J., Feng, G., Fu, J., Bao, J., Zhang, Z. and He, L.: An association study of the N-methyl-D-aspartate receptor NR1 subunit gene (GRIN1) and NR2B subunit gene (GRIN2B) in schizophrenia with universal DNA microarray, *Eur. J. Hum. Genet.*, Vol.13, pp.807–814 (2005).
12) Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F. and Moore, J.H.: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer, *Am. J. Hum. Genet.*, Vol.69, pp.138–147 (2001).
13) Bellman, R.: *Adaptive Control Processes*, Princeton University Press, Princeton, New Jersey (1961).
14) Sokal, R.R. and Rohlf, F.J.: *The principles and practice of statistics in biological research*, Biometry, WH Freeman and company, New York, 3rd edition (1995).
15) Listgarten, J., Damaraju, S., Poulin, B., Cook, L., Dufour, J., Driga, A., Mackey, J., Wishart, D., Greiner, R. and Zanke, B.: Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms, *Clin. Cancer Res.*, Vol.10, pp.2725–2737 (2004).
16) Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T.R. and Mesirov, J.P.: Estimating dataset size requirements for classifying DNA microarray data, *J. Comput. Biol.*, Vol.10, pp.119–142 (2003).
17) Olshen, A.B. and Jain, A.N.: Deriving quantitative conclusions from microarray expression data, *Bioinformatics*, Vol.18, pp.961–970 (2002).
18) Freund, Y. and Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting, *J. Comput. Sys. Sci.*, Vol.55, pp.119–139 (1997).
19) Yamada, Y., Ichihara, S., Izawa, H., Tanaka, M. and Yokota, M.: Association of a G994 → T (Val279 → Phe) polymorphism of the plasma platelet-activating factor acetylhydrolase gene with myocardial damage in Japanese patients with nonfamilial hypertrophic cardiomyopathy, *J. Hum. Genet.*, Vol.46, pp.436–441 (2001).

Appendix

A.1 Abbreviation Index

gene symbol	*gene*
AGT	*Angiotensinogen*
APOCIII	*Apolipoprotein C-III*
APOE	*Apolipoprotein E*
CCR2	*CC chemokine receptor 2*
CD14	*CD14 receptor*
CX37	*Connexin 37*
ET1	*Endothelin-1*
FABP2	*Fatty acid binding protein 2*
GNB3	*G protein β3 subunit*
GP1A	*Glycoprotein Ia*
GP1BA	*Glycoprotein Iba*
IL10	*Interleukin-10*
IL6	*Interleukin-6*
IRS1	*Insulin receptor substrate-1*
MMP3	*Stromelysin-1*
NOS3	*Endothelial nitric oxide synthase*
P22	*p22phox*
PAI1	*Plasminogen-activator inhibitor type 1*
PLA2G7	*Platelet-activating factor acetylhydrolase*
PON1	*Paraoxonase*
SELE	*E-selectin*
TAP	*ATP-binding cassette transporter*
TGFB1	*Transforming grpp. owth factor β1*
THBD	*Thrombomodulin*
THBS4	*Thrombospondin 4*
THPO	*Thrombopoietin*
TNFA	*Tumor necrosis factor α*

(Received May 22, 2006)
(Accepted July 10, 2006)

(Communicated by *Susumu Goto*)

Yasuyuki Tomita was born in 1980. He received his B.E. and M.E. degrees from Nagoya University in 2003 and 2005 respectively. He is currently a Ph.D. candidate of the Graduate School of Engineering, Nagoya University. His research interests are biostatistics and bioinformatics.

Hiroyuki Asano was born in 1968. He received his M.D. degree from Saga Medical School, Japan, in 1994. He is a student of the Department of Cardiology, Internal Medicine, Program in Integrated Molecular Medicine, Nagoya University, Graduate School of Medicine.

Hideo Izawa was born in 1964. He received his M.D. degree from Nagoya University, Japan in 1989, and Ph.D. degree from Nagoya University in 1998, respectively. He is an Assistant Professor of the Department of Cardiology, Nagoya University, Graduate School of Medicine.

Mitsuhiro Yokota received his M.D. degree from Nagoya University School of Medicine, Japan in 1969, and Ph.D. degree from Nagoya University School of Medicine in 1979, respectively. Since 2004 he had been a Professor of Department of Cardiovascular Genome Science, Nagoya University School of Medicine. He is currently a Professor of the Department of Genome Science, Aichi-Gakuin University, School of Dentistry.

Takeshi Kobayashi was born in 1941. He received his Ph.D. degree from Nagoya University, Japan in 1969. Since 1982 he had been a Professor of Nagoya University, Graduate School of Engineering. He is currently a Professor of the Department of Applied Biological Science, Chubu University.

Hiroyuki Honda was born in 1961. He received his Ph.D. degree from Nagoya University, Japan in 1988. Since 2004 he has been a Professor of Nagoya University, Graduate School of Engineering. One of his reserch interests is knowledge information processing.