Classification method for prediction of multifactorial disease development using interaction between genetic and environmental factors

Yasuyuki Tomita[1], Mitsuhiro Yokota[2], Hiroyuki Honda*[1]
[1]*Department of Biotechnology, School of Engineering, Nagoya University*
[2]*Department of Cardiovascular Genome Science, School of Medicine, Nagoya University*
**e-mail: honda@nubio.nagoya-u.ac.jp*

Abstract

Multifactorial disease such as life style related diseases, for example, cancer, diabetes mellitus, myocardial infarction (MI) and others, is thought to be caused by complex interactions between polygenic basis and various environmental factors. In this study, we used 22 polymorphisms on 16 candidate genes that have been characterized and potentially associated with MI in terms of biological function and 6 environmental factors. To predict development for MI and classify the subjects into personally optimum development patterns, we extracted risk factor candidates (RFCs) composed of state which is a derivative form of polymorphisms and environmental factors using statistical test and selected risk factors from RFCs using Criterion of Detecting Personal Group (CDPG) defined in this study. We could predict development of blinded data simulated as unknown their development more than 80% accuracy and identify their causal factors using CDPG.

1. Introduction

Recently, genetic linkage and association studies have already identified several candidate genes that may predispose to MI [1]. Thus genetic factors may be necessary for development of the disease, but the disease would not be manifested without the presence of an environmental risk factor [1]. The methods to detect interaction between gene and environment, or gene and gene, and predict development of multifactorial disease with high accuracy have been scarcely proposed as attractive and convenient tools with sufficient performance. In the present study, (1) analysis of exhaustive combination consisting of up to 3 factors was performed and risk factor candidates (RFCs) were extracted using binomial test and random permutation test. (2) To classify the blinded data into personally optimum development patterns, Criterion of Detecting Personal Group (CDPG) was newly defined in the present study, and selection of the smallest number of risk factors from RFCs and prediction of their development were performed.

2. Methods

2.1 Extraction of risk factor candidates

We used 22 polymorphisms on 16 candidate genes and 6 environmental factors (smoking, obesity, hypertension, diabetes mellitus, hypercholesterolemia and hyperuricemia) with thousands of subjects (subjects with MI and no symptoms of MI). The data was divided into 2 data sets (modeling data; MD and blinded data; BD). We performed exhaustive combination analysis using MD and assumed the appearance of case and control subjects belonging to a certain rule l (Figure. 1) as a series of *Bernoulli trials*, where two possible outcomes are case and control subjects and they occur with the probabilities of $N_{case,l}/(N_{case,l}+N_{control,l})$ and $N_{control,l}/(N_{case,l}+N_{control,l})$, respectively. The number of tirals (n) is the sum of the observed number for $N_{case,l}$ and $N_{control,l}$. In this case, binomial distribution of case subjects is as follows. N_{case} and $N_{control}$ represent the number of the whole case and control subjects analyzed in the combination. The probability p represents $N_{case}/(N_{case}+N_{control})$.

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

The null hypothesis ($N_{case,l}/N_{case} \leq N_{control,l}/N_{control}$) is tested by computing the sum (P-value) of all $f(N_{case,l})$ which are equal to or smaller than that for the observed value of $N_{case,l}$ (one-tailed test). In order to extract risk factor candidates (RFCs), statistical significance of rule in each combination assigned to the P-value by modeling the null distribution which was the lowest P-value in each combination with random permutation test [2]. In the present study, RFCs were inferred at the P-value level using this distribution calculated with random permutation test less than 0.01.

Rule Table			Polymorphism A	
			AA	Aa+aa
Polymorphism B	BB+Bb	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	$N_{case,7}/N_{control,7}$	$N_{case,8}/N_{control,8}$

environmental factor

Figure 1. The rule table using combination between two polymorphisms and one environmental factor.

2.4 Selection of risk factors from RFCs

We suggested Criterion of Detecting Personal Group (CDPG) to select the smallest number of risk factors in order to classify the BD into personally optimum development patterns and predict their development. The selection of m'th risk factor is carried out in order to maximize the following index I.

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

$N^{(m)}_{RFC,case}$ and $N^{(m)}_{RFC,control}$ represents the number of case and control subjects who have more than one RFC in selecting m'th risk factor. N_{case} and $N_{control}$ represent the number of case and control subjects in MD. In BD, if the subject had one risk factor at least, the prediction was case and if the subject had no risk factor, the prediction was control.

3. Results and Discussion

The relationship between cover rate (subjects with the RFC/whole subjects) and case rate (case subjects with the RFC/ subjects with the RFC) of each extracted RFCs shown in Figure 2. This tendency was found in BD. In 1,165 RFCs, we selected 26 risk factors which classified 2,572 subjects in MD into 26 development patterns and decided risk factors individually. The development in 80.2% (142/177) of the case subjects in BD were predicted correctly using selected 26 risk factors. In addition, we examined the relationship between the number of subjects and the number of risk factors (NRF) which they had in 26 selected risk factors (the result of BD; 286 subjects shown in Figure 3). It was found that risk rate was higher with increasing NRF. Consequently, this method is effective for preventive medicine of multifactorial disease using polymorphism and environmental factor data.

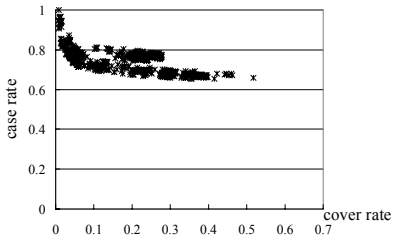

Figure 2. Cover and case rate of 1,165 RFCs.

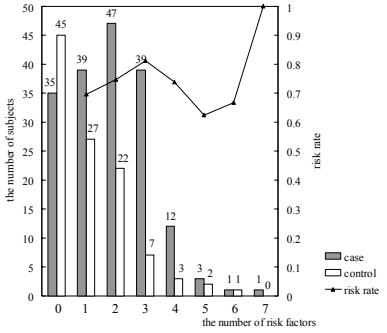

Figure 3. The risk rate cut with NRF in BD.

4. References

[1] Y. Yamada, H. Izawa, S. Ichihara, F. Takatsu, H. Ishihara, H. Hirayama, T. Sone, M. Tanaka and M. Yokota, "Prediction of the risk of myocardial infarction from polymorphisms in candidate genes", *N. Engl. J. Med.*, vol.347, no.24, Dec.2002, pp.1916-1923.

[2] S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T.R. Golub and J.P. Mesirov, "Estimating dataset size requirements for classifying DNA microarray data" *J. Comput. Biol.*, vol.10, no.2, Apr.2003, pp.119-142.