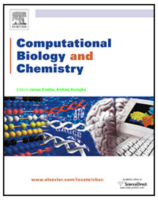
Computational Biology and Chemistry 65 (2016) 193–202

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem



A novel fuzzy set based multifactor dimensionality reduction method for detecting gene–gene interaction

Hye-Young Jung[a,1], Sangseob Leem[b,1], Sungyoung Lee[c], Taesung Park[b,c,*]

[a] Faculty of Liberal Education, Seoul National University, Seoul, 08826, South Korea
[b] Department of Statistics, Seoul National University, Seoul, 08826, South Korea
[c] Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul, 08826, South Korea

ARTICLE INFO

Article history:
Received 28 August 2016
Accepted 7 September 2016
Available online 29 September 2016

Keywords:
Gene-gene interaction
Multifactor dimensionality reduction
Fuzzy classifier
Fuzzy set theory
Uncertainty

ABSTRACT

Background: Gene-gene interaction (GGI) is one of the most popular approaches for finding the missing heritability of common complex traits in genetic association studies. The multifactor dimensionality reduction (MDR) method has been widely studied for detecting GGIs. In order to identify the best interaction model associated with disease susceptibility, MDR compares all possible genotype combinations in terms of their predictability of disease status from a simple binary high(H) and low (L) risk classification. However, this simple binary classification does not reflect the uncertainty of H/L classification.

Methods: We regard classifying H/L as equivalent to defining the degree of membership of two risk groups H/L. By adopting the fuzzy set theory, we propose Fuzzy MDR which takes into account the uncertainty of H/L classification. Fuzzy MDR allows the possibility of partial membership of H/L through a membership function which transforms the degree of uncertainty into a [0,1] scale. The best genotype combinations can be selected which maximizes a new fuzzy set based accuracy measure.

Results: Two simulation studies are conducted to compare the power of the proposed Fuzzy MDR with that of MDR. Our results show that Fuzzy MDR has higher power than MDR. We illustrate the proposed Fuzzy MDR by analysing bipolar disorder (BD) trait of the WTCCC dataset to detect GGI associated with BD.

Conclusions: We propose a novel Fuzzy MDR method to detect gene–gene interaction by taking into account the uncertainly of H/L classification and show that it has higher power than MDR. Fuzzy MDR can be easily extended to handle continuous phenotypes as well. The program written in R for the proposed Fuzzy MDR is available at https://statgen.snu.ac.kr/software/FuzzyMDR.

1. Background

Gene-gene interactions (GGIs) are one of the most important contributors to the variation of complex traits because many biological phenotypes result from the complex interplay of multiple genes and environmental factors. Detection of GGI or epistasis has been recognized as one of the most effective remedies for the problem of missing heritability in genome-wide association studies (GWAS) (Mackay, 2014; Eichler et al., 2010). Many efficient approaches have been proposed for GGI analysis including model-based approaches and model-free approaches. Model-based approaches assume certain statistical models between genotype and phenotype (Wu et al., 2009; Yang et al., 2009; Wan et al., 2010a; Park and Hastie, 2008; Zhang and Liu, 2007), while the model-free approaches often have no prior assumption about the model and data (Ritchie et al., 2001; Zhang et al., 2010; Dong et al., 2008; Li et al., 2014).

Among the model-free approaches, the Multifactor Dimensionality Reduction (MDR) approach (Ritchie et al., 2001) is a very popular non-parametric, combinatory approach. It reduces the number of dimensions by converting a high-dimensional multi-locus model into a one-dimensional model. MDR reduces multiple genotype combinations into two groups—high (H) and low (L) risk groups.

There are a number of extensions of MDR: quantitative MDR (QMDR) for a quantitative trait (Gui et al., 2013), generalized MDR (GMDR) for both quantitative and binary traits (Lou et al., 2007), Surv-MDR and Cox-MDR for survival data (Lee et al., 2011, 2012), FAM-MDR for family data (Cattaert et al., 2010; Lou et al., 2008),

[*] Corresponding author.
E-mail addresses: hyjunglove@snu.ac.kr (H.-Y. Jung), leemss@snu.ac.kr (S. Leem), biznok@gmail.com (S. Lee), tspark@stats.snu.ac.kr (T. Park).
[1] Joint first authors.

http://dx.doi.org/10.1016/j.compbiolchem.2016.09.006
1476-9271/

GEE-MDR and Multi-MDR for multivariate traits (Choi and Park, 2013; Yu et al., 2015). These extensions of MDR have been shown to have great power in broad applications of identifying high-order interactions.

However, one of the shortcomings of MDR lies in the uncertainty of simple binary high(H)/low(L) classification. In MDR analysis, binary classification compares the conditional odds of case and control given a genotype combination to the unconditional odds of the total numbers of cases and controls. Although binary classification provides a straightforward interpretation of results, it suffers from a loss of information. For example, assume that there are the same numbers of cases and controls. Then, the unconditional odds of cases and controls is one. Assume that there are two single nucleotide polymorphisms (SNPs). These two SNPs generate nine possible genotype combinations. Suppose that one genotype combination is classified as H with a conditional odds of 10 and another as H with a conditional odds of 1.5. Though the difference between these odds is quite large, the original MDR does not distinguish the two genotype combinations. Both genotype combinations are classified as H. This motivated our earlier development of OR MDR (Chung et al., 2007), which provides the estimated odds ratio (OR) as a quantitative measure of disease risk for each multi-locus genotype. Our other work, called wBA MDR hereafter, used the ORs as weights for computing the weighted balanced accuracy (wBA) in order to take into account these differences (Namkung et al., 2009).

In association with these works, we now propose a novel MDR extension with a totally different paradigm. We regard classifying H/L as equivalent to defining a degree of membership of two risk groups H and L. By adopting the fuzzy set theory, we propose Fuzzy MDR which takes into account the uncertainty of binary classification by allowing the possibility of partial membership of H/L. For example, for two genotype combinations with conditional odds 10 and 1.5, Fuzzy MDR allows the assignment of different membership functions to the two genotype combinations. That is, Fuzzy MDR assigns the first genotype combination to (H, L) with a membership value (1.0, 0), while it assigns the second combination to (H, L) with a different a membership value (0.6, 0.4). The membership value (μ_H, μ_L) with a constraint $\mu_H + \mu_L = 1$ allows partial membership and transforms the degree of uncertainty into a [0,1] scale. The best genotype combinations can then be selected using a new fuzzy set based accuracy measure representing the degree of uncertainty. A more systematic illustration through motivating examples is given in the next section (Fig. 1).

The next section describes the Fuzzy MDR method with an introduction to the fuzzy set theory and membership function. To evaluate the power of Fuzzy MDR and compare its power with that of MDR, extensive simulation studies are conducted with two datasets: one without marginal effects and the other with marginal effects. Finally, the proposed Fuzzy MDR method is applied to the bipolar disorder (BD) trait from Wellcome Trust Case Control Consortium (WTCCC) data to illustrate its performance (Consortium WTCC, 2007).

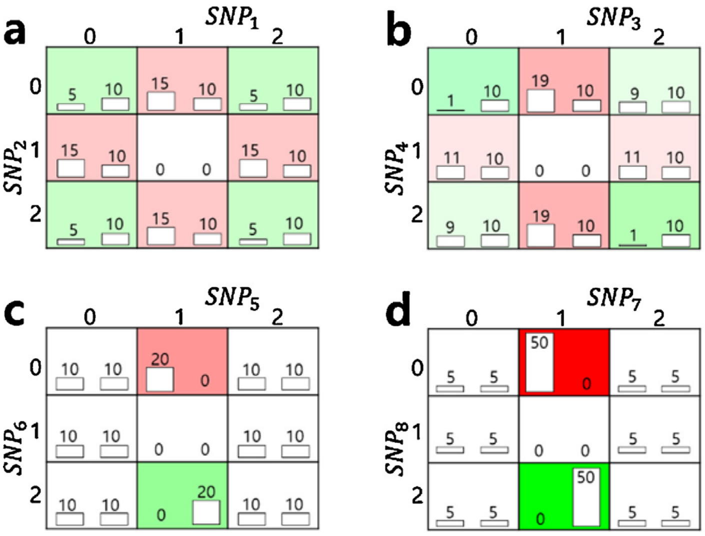
2. Methods

2.1. Motivation

We first introduce a motivating example for Fuzzy MDR. Suppose that there are two SNPs of interest. From the two SNPs of interest, we can generate nine genotype combinations that can be summarized in a 3 × 3 contingency table. Fig. 1 shows four examples of 3 × 3 contingency tables. Each cell representing one genotype combination contains two bars: the first one for the cases and the second one for the controls. The red color represents the cells classified as a high (H) risk group and the green color represents those classified as a low (L) risk group. The total numbers of cases and controls are assumed to be 80 each. Let n_{i0} and n_{i1} represent the frequencies of the i-th genotype combination for the controls and cases, respectively. Here, i enumerates all possible genotype combinations: for the two SNPs, $i = \{1, \cdots, 9\}$ and for p SNPs, $i = \{1, \cdots, 3^p\}$. The first two Fig. 1a and b have four H cells and four L cells. Both figures have the same values for the following sums:

$$\sum\nolimits_{i \in H}(n_{i1} - n_{i0}) \text{ and } \sum\nolimits_{i \in L}(n_{i1} - n_{i0}).$$

Note that each H (L) cell in Fig. 1a has the same difference of $n_{i1} - n_{i0}$. On the other hand, these differences vary across cells in Fig. 1b. Therefore, the two SNPs in Fig. 1b are more strongly associated with the disease status than the two SNPs in Fig. 1a, which can be confirmed by statistical tests. The chi-square statistics for Fig. 1a and b are 10.667 (p-value = 0.221) and



method	a	b
Chi-square statistic (p-value)	10.667 (0.221)	20.514 (0.009)
Balanced accuracy of MDR	0.625	0.625
Balanced accuracy of wBA MDR (α = 0.25)	0.629	0.667
Balanced accuracy of fuzzy MDR (linear, OR = 8)	0.531	0.570

method	c	d
Chi-square statistic (p-value)	17.501 (0.025)	35.029 (2.6E-5)
Balanced accuracy of MDR	0.625	0.813
Balanced accuracy of wBA MDR (α = 0.25)	1.000	1.000
Balanced accuracy of fuzzy MDR (linear, OR = 8)	0.625	0.813

Fig. 1. Examples representing drawback of MDR. Bars in each cell represent the number of cases and controls. The red color represents the cells classified as high (H) risk and the green color does those classified as low (L) risk. The intensity of background color represents the degree of membership function. The white background color means the tied cells with equal number of cases and controls. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)