BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 17 2010, pages 2198–2199
doi:10.1093/bioinformatics/btq352

Genetics and population analysis

Advance Access publication July 1, 2010

mbmdr: an R package for exploring gene–gene interactions associated with binary or quantitative traits

M. Luz Calle[1,*], Víctor Urrea[1], Núria Malats[2] and Kristel Van Steen[3]

[1]Department of Systems Biology, Universitat de Vic, Vic, [2]Centro Nacional de Investigaciones Oncológicas, Madrid, Spain and [3]Department of Electrical Engineering and Computer Science, University of Liège, Liège, Belgium

Associate Editor: Martin Bishop

ABSTRACT

Summary: We describe mbmdr, an R package for implementing the model-based multifactor dimensionality reduction (MB-MDR) method. MB-MDR has been proposed by Calle *et al*. as a dimension reduction method for exploring gene–gene interactions in case-control association studies. It is an extension of the popular multifactor dimensionality reduction (MDR) method of Ritchie *et al*. allowing a more flexible definition of risk cells. In MB-MDR, risk categories are defined using a regression model which allows adjustment for covariates and main effects and, in addition to the classical low risk and high risk categories, MB-MDR considers a third category of indeterminate or not informative cells. An important improvement added to the current mbmdr algorithm with respect to the original MB-MDR formulation in Calle *et al*. and also to the classical MDR approach, is the extension of the methodology to different outcome types. While MB-MDR was initially proposed for binary traits in the context of case-control studies, the mbmdr package provides options to analyze both binary or quantitative traits for unrelated individuals.

Availability: http://cran.r-project.org/

Contact: malu.calle@uvic.cat

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 12, 2010; revised on June 7, 2010; accepted on June 29, 2010

1 INTRODUCTION

Model-based multifactor dimensionality reduction (MB-MDR), has been proposed by Calle *et al*. (2008) as a dimension reduction method for exploring gene–gene interactions in case-control association studies. MB-MDR extends the multifactor dimensionality reduction (MDR) method of Ritchie *et al*. (2001) in several ways. Like MDR, the MB-MDR method merges multi-locus genotypes into a one-dimensional construct, but the way the genotype cells are merged differs. In MB-MDR, an additional 'no-evidence category' allows a more accurate categorization into high level, low level and 'indeterminate' or 'non-informative' cells.

Other extensions to improve the performance and applicability of MDR have been proposed, including the odds ratio-based MDR (OR-MDR) method (Chung *et al*., 2007) and the generalized MDR GMDR; (Lou *et al*., 2007). Unlike these MDR extensions, MB-MDR does not involve division of the data into training and learning sets and therefore does not select the best interaction model via prediction accuracy and cross-validation consistency measures. Instead, MB-MDR measures the association between multi-locus genotypes and the phenotype and provides a set of statistically significant interactions instead of a single best model. Significance is assessed through a permutation test. This strategy has proven to be more powerful than MDR in the presence of genetic heterogeneity (Cattaert *et al*., submitted for publication).

An additional feature of MB-MDR is its flexibility for dealing with different kind of phenotypes by changing the link function of the regression analysis. MB-MDR was initially proposed for binary traits in the context of case-control studies (Calle *et al*., 2008), but the mbmdr package presented here has been implemented to allow for quantitative traits.

MB-MDR has been successfully applied to a set of 404 single nucleotide polymorphisms (SNPs) in 110 inflammation-related genes as a preliminary phase to identify epistatic effects in the Spanish Bladder Cancer/EPICURO Study (Calle *et al*., 2008). This is a case-control study conducted in 18 Spanish centers during 1998–2001 that included 1157 cases and 1157 controls. Unlike MDR, this approach allowed to adjust the analysis for the main bladder cancer risk factors (smoking status, age, gender and region). MB-MDR identified one second-order interaction and five three-way interactions with $P < 10^{-5}$. While some interactions seem to be biologically sound, others were unexpected and need to be replicated in independent series before validating them in the lab.

MB-MDR has also been applied to a study on eczema risk (Mahachie John *et al*., 2009) where the phenotype was the allergen-specific IgE levels. This application is specially relevant since, involving a continuous response, the classical MDR approach could not be used. Furthermore, the flexible model-based approach allowed adjustment for well-known established risk factors. A significant epistatic effect of two variants in FCER1A on eczema risk was detected.

2 MB-MDR OVERVIEW AND IMPLEMENTATION

MB-MDR approach consists of three steps that are briefly described next. A more detailed description can be found in Calle *et al*. (2007, 2008). mbmdr package contains two main functions: `mbmdr` that performs Steps 1 and 2 and `mbmdr.PermTest` that performs Step 3, the permutation approach for significance assessment.

We describe the results of applying the mbmdr package to an example dataset included in the package. The detailed code of this implementation is provided in the Supplementary Material. The dataset `simSNPS` contains a simulated dataset from epistatic

*To whom correspondence should be addressed.

model 1 described in Ritchie *et al*. (2003). It contains 10 SNPs for 400 individuals, 200 cases and 200 controls, and a covariate associated with response. Only the second-order interaction between SNP1 and SNP2 is functional.

Step 1: The first step performs an association test for each multilocus cell with the phenotype: a logistic regression for binary traits, a linear regression for quantitative traits or any other specified link function. Each genotype cell is then assigned to one of three categories, high risk (*H*), low risk (*L*) or no evidence (0), as a result of the association test with a liberal threshold of 0.1 for assigning the genotype to a risk category, *H* or *L*. The following lines are the output of MB-MDR step1 for interaction between SNP1 and SNP2, using function `mbmdr` and adjusting for covariate X. It provides the results of the different regressions applied to each multilocus genotype:

```
SNP1 SNP2 cases controls beta  pval   cat
  0    0     0     9 -0.51 8.3e-03  L
  1    0    49    19  0.28 1.9e-04  H
  2    0     0    15 -0.52 6.5e-04  L
  0    1    50    19  0.16 3.0e-02  H
  1    1     0    57 -0.51 4.1e-10  L
  2    1    43    30  0.11 1.1e-01  0
  0    2     0    14 -0.21 1.7e-01  0
  1    2    58    26  0.18 8.6e-03  H
  2    2     0    11 -0.50 4.3e-03  L
```

In this example, the interaction between SNP1 and SNP2 is highly predictive and provides perfect classification of cases and controls in some cells. This perfect classification is known as the separation problem (Heinze and Schemper, 2002) and provides parameter estimates extremely inaccurate (equal to infinite). To avoid this and provide accurate estimates, the `mbmdr` calls the R package `logistf` that implements a penalized likelihood.

Step 2: After merging the multilocus genotypes of the same risk category, Step 2 performs two new association tests, one for each risk category, high and low, on the outcome variable. In each test, the group of interest is compared with the other two groups using a regression model. This second step provides a Wald statistic, WH, for the high risk category H versus {L,0} and a Wald statistic, WL, for the low risk category L versus {H,0}. The test statistic for the epistatic effect will be based on the maximum between WH and WL.

The output of MB-MDR Step 2 provides the number of multilocus genotypes classified as High risk (NH), the Wald statistic for the High risk category (WH) and the corresponding unadjusted *P*-value (PH). The same information is provided for the Low risk category (NL, WL, PL). The minimum between PH and PL is given in the last column (MIN.P):

```
SNP1  SNP2  NH   WH    PH     NL   WL    PL   MIN.P
SNP1  SNP2   3 76.62 2.0e-18   4 0.00  0.97 2.0e-18
```

Step 3: Explores significance of the specified models through a permutation test on the maximum Wald statistic and is implemented by the function `mbmdr.PermTest`. The procedure also can provide the confidence intervals of the permuted *P*-value (Nettleton and Doerge, 2000) at a given significance level. The output of the permutation test provides again the relevant information of the risk class (NH, WH, NL, WL), the maximum between the two Wald statistics (W_{max} = max(WH, WL)) and the adjusted permutation *P*-value obtained from the permutation distribution of Wmax:

```
SNP1  SNP2   NH  WH   NL   WL  Wmax  Perm.P
SNP1  SNP2   3  76.62  4  0.97 76.62  <1e-04
```

Instead of exploring each interaction separately, the default call of `mbmdr` function is to explore all possible interactions of a given order at a time. However, since the permutational approach is very time consuming and accurate *P*-values require a large number of permutations, we suggest to assess their significance in a sequential way that discards from further exploration those models with no signal of association (Supplementary Material).

Computer details and time: the full process of exploring all possible 45 second-order interactions has been run under R software, as a single thread on a server environment running a 64 bits linux distribution with Quad Core Intel Xeon processor at 2.5 GHz, 12 MB of L2 memory, 1333 MHz of FSB and 8 GB of RAM. The running time for Steps 1 and 2 was 7 s and for Step 3 (permutational significance) with 10 000 permutation was 54 min and 25 s. In studies involving a larger number of SNPs, computational time is an issue and, in some cases, a full exploration of significance will be unfeasible (Supplementary Material). In this case preselection of SNPs will be required, as in Calle *et al*. (2008) where the MB-MDR methodology is proposed in combination with preselection based on the observed synergy between SNPs.

Funding: The Ministerio de Educación y Ciencia (Spain) (MTM2008-06747-C02-02); La Marató de TV3 Foundation (Grant 050831); Generalitat de Catalunya (2009SGR-581).

Conflict of Interest: none declared.

REFERENCES

Calle,M.L. *et al*. (2007) MB-MDR: model-based multifactor dimensionality reduction for detecting interactions in high-dimensional genomic data. *Technical Report n.24*, Universitat de Vic, http://www.recercat.net/handle/2072/5001 (last accessed date July 14, 2010).

Calle,M.L. *et al*. (2008) Improving strategies for detecting genetic patterns of disease susceptibility in association studies. *Stat. Med.*, **27**, 6532–6546.

Chung,Y. *et al*. (2007) Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. *Bioinformatics*, **23**, 71–76.

Heinze,G. and Schemper,M. (2002) A solution to the problem of separation in logistic regression. *Stat. Med.*, **21**, 2409–2419.

Lou,X. *et al*. (2007) A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *Am. J. Hum. Genet.*, **80**, 1125–1137.

Mahachie John,J.M. *et al*. (2009) Analysis of the high affinity IgE receptor genes reveals epistatic effects of FCER1A variants on eczema risk. *Allergy*, **65**, 875–882.

Nettleton,D. and Doerge,R.W. (2000) Accounting for variability in the use of permutation testing to detect quantitative trait loci. *Biometrics*, **56**, 52–58.

Ritchie,M.D. *et al*. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, **69**, 138–147.

Ritchie,M.D. *et al*. (2003) Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet. Epidemiol.*, **24**, 150–157.