Bioinformatics, 31(11), 2015, 1842–1844
doi: 10.1093/bioinformatics/btv021
Advance Access Publication Date: 20 January 2015
Applications Note

OXFORD

Gene expression

dslice: an R package for nonparametric testing of associations with application in QTL and gene set analysis

Chao Ye[1], Bo Jiang[2], Xuegong Zhang[1,3,*] and Jun S. Liu[2,4,*]

[1]MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST, Department of Automation, Tsinghua University, Beijing 100084, China, [2]Department of Statistics, Harvard University, Cambridge, MA 02138, USA, [3]School of Life Sciences, Tsinghua University, Beijing 100084, China and [4]Center of Statistics, Tsinghua University, Beijing 100084, China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on July 3, 2014; revised on December 28, 2014; accepted on January 12, 2015

Abstract

Summary: Many statistical problems in bioinformatics and genetics can be formulated as the testing of associations between a categorical variable and a continuous variable. A dynamic slicing method was proposed for non-parametric dependence testing, which has been demonstrated to have higher powers compared with traditional methods such as Kolmogorov–Smirnov test. We introduce an R package *dslice* to facilitate the use of dynamic slicing method in bioinformatic applications such as quantitative trait loci study and gene set enrichment analysis.

Availability and implementation: *dslice* is implemented in Rcpp and available in the Comprehensive R Archive Network. The package is distributed under the GNU General Public License (version 2 or later).

Contact: zhangxg@tsinghua.edu.cn or jliu@stat.harvard.edu.

1 Introduction

Often in biological studies, we need to test whether the underlying distributions of two or more populations are different from each other on the basis of independent samples from these populations. The *K*-sample test problem is equivalent to testing whether the value of an observation is independent of its label indicator (i.e. which sample it comes from). Traditional methods such as the *t*-test and the rank-sum test can only test differences in means or medians. On the other hand, classic omnibus testing methods such as Kolmogorov–Smirnov test have limited power in detecting differences between distributions of different types. Jiang *et al.* (2014) proposed an omnibus *K*-sample testing method based on regularized likelihood-ratio test and dynamic slicing (here we refer to the test statistic as 'DS-statistic') and demonstrated its statistical power compared with some existing well-known methods through extensive simulation studies. Here, we introduce an R package *dslice* that implements this novel dynamic slicing ('DS') method for nonparametric tests and makes it a versatile statistical tool for applications in quantitative trait loci (QTL) study and gene set enrichment analysis (GSEA).

2 Implementation

Core functions for testing dependence in *dslice* are implemented in the Cpp language and are integrated in R through the Rcpp package (Eddelbuettel and François, 2011). R package *dslice* contains functions for omnibus *K*-sample hypothesis testing (*ds_test*), illustrations of optimal slicing scheme (*slice_show*) and gene set analysis (*ds_gsa*) on dataset downloaded from GSEA website (http://www.broadinstitute.org/gsea/index.jsp).

3 Examples

3.1 QTL study

The QTLs analysis attempts to detect relationship between genetic variation (single-nucleotide polymorphism) and continuous variable

Table 1. Identified QTLs associated with mouse femur and vertebra length by the DS method

Phenotype	QTL	Location	DS-statistic	FDR*
Femur length	rs3091203	Chr13:22	18.34	$<1\times10^{-6}$
	D2Mit285	Chr2:159	13.79	$<1\times10^{-6}$
	D9Mit104.1	Chr9:64	5.87	0.0055
	rs13482609	Chr15:64	3.47	0.0257
	D5Mit25.1	Chr5:111	3.4	0.0278
	rs3657845	Chr17:18	2.79	0.0393
Vertebra length	rs4222738	Chr1:161	23.11	$<1\times10^{-6}$
	D1Mit105	Chr1:163	16.58	$<1\times10^{-6}$
	D2Mit58	Chr2:110	8.71	2.6×10^{-4}
	rs3089785	Chr16:39	5.01	0.0044
	D7Mit76	Chr7:17	4.99	0.0048
	rs4231015	Chr15:100	4.2	0.0086
	D5Mit48	Chr5:9	4.16	0.0086
	rs13477864	Chr4:99	3.49	0.0153
	rs13480044	Chr8:129	3.07	0.0205
	rs13477990	Chr4:133	2.72	0.0256
	rs3023444	Chr17:38	2.53	0.0282
	rs4229231	Chr12:5	2.08	0.0385
	rs3091203	Chr13:22	1.82	0.0438
	D19Mit41	Chr19:20	1.72	0.0496
	D14Mit170	Chr14:105	1.63	0.0494

Genomic location is in chromosome: Mb format.
FDR*, FDR-adjusted *P* values.

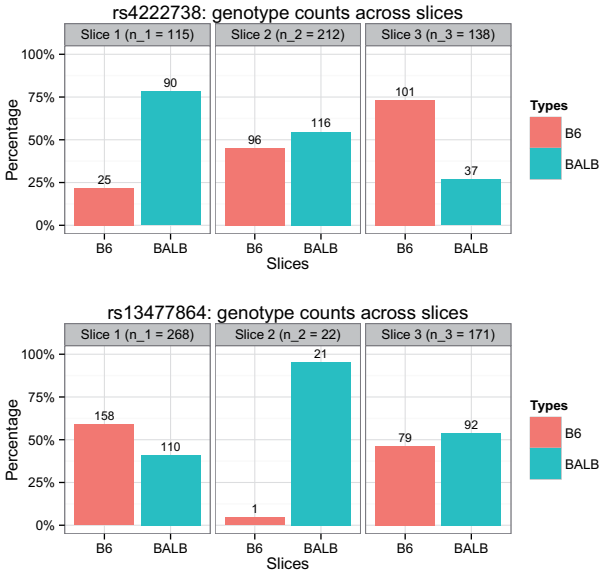

Fig. 1. Illustration of optimal slicing schemes for mouse vertebra length QTLs *rs4222738* and *rs13477864*. Each panel shows the relative proportion of genotypes in the corresponding slice

Table 2. Significant gene sets detected by the DS method

Gene set	Size	DS-statistic	FDR*
P53 pathway	16	10.42	$<1\times10^{-6}$
P53 hypoxia pathway	20	9.69	$<1\times10^{-6}$
P53_signalling	87	5.095	$<1\times10^{-6}$
Radiation_sensitivity	26	8.846	3.1×10^{-4}
P53_UP	40	11.2	0.0022
hsp27 pathway	15	0.514	0.0456

FDR*, FDR-adjusted *P* values.

of phenotypes (human height or gene expression levels). Traditional analyzing methods are mainly based on the researcher's foresight about the agnostic mechanism of QTL, such as linear marginal effect or interaction effect (Aschard *et al.*, 2013). Dynamic slicing method views the QTL problem as dependence test between a categorical variable (genotype) and a continuous variable (phenotype). It has the advantage of revealing nonlinear associations free of assumptions on underlying mechanism.

In view of power study on DS method in detecting non-linear effects (Jiang *et al.*, 2014), we apply the *ds_test* function in *dslice* package on a mouse QTL dataset by Burke *et al.* (2012) with 558 binary genetic markers and two phenotypes. We removed observations with missing phenotype values and set the penalty parameter $\lambda = 1.0$ in *ds_test*. Then, we randomly shuffled the phenotypes of individuals to generate empirical *P* values of DS-statistic. By controlling a false discovery rate (FDR, Benjamini and Hochberg, 1995) of 0.05, we identified 6 QTLs and 15 QTLs associated with mouse femur length and vertebra length, respectively (Table 1). Figure 1 illustrates the optimal slicing schemes generated by *dslice* for mouse vertebra length QTLs *rs4222738* and *rs13477864*. We can see that non-linear associations like the example of QTL *rs13477864* can be well detected by *dslice*.

3.2 Gene Set Analysis

Subramanian *et al.* (2005) introduced GSEA to the aggregate effect of genes in unit of 'gene set'. Specifically, GSEA attempts to determine whether the distribution of biological phenotypes are different between genes in a gene set and the other genes, which can be formulated as a non-parametric two-sample testing problem. However, Goeman and Bühlmann (2007) discussed two different null hypotheses in gene set analysis: competitive null, which is a two-sample test comparing genes in the gene set and genes not in the set, and self-contained null, which is a one-sample test of differential expression on a set of genes. Here, we focus on the application of dynamic slicing method on a well-studied dataset P53 NCI-60 (Ackermann and Strimmer, 2009; Efron and Tibshirani, 2007; Subramanian *et al.*, 2005) in testing competitive null but note that our method can also be used in one-sample test for self-contained null (see function *ds_* 1 in *dslice* package for details).

This dataset assays 10 100 gene expression levels and consists of 17 normal samples and 33 samples with mutated p53. The C2 gene set contains 308 predefined functional gene sets (with gene set size between 15 and 500). The dataset is available on the GSEA website. We set the penalty parameter $\lambda = 1.0$ and obtained the empirical *P* values by randomly assigning sample labels. Table 2 lists significant gene sets reported by DS under a FDR cutoff of 0.05. In comparison, the GSA method by Efron and Tibshirani (2007) missed the gene set *p53_signalling* under the same cutoff, and all the significant gene sets reported by GSA have FDR values larger than 0.01.

4 Discussion

Testing of associations between a categorical variable and a continuous response is a frequently encountered statistical problem in bioinformatics. R package *dslice* implements a recently proposed nonparametric dependency detection method. Although we use QTL and gene set analysis as examples to demonstrate the use of *dslice*, it can be applied to other biological problems such as protein

binding inference from nucleosome occupancy (Meyer *et al*., 2011) and binding profile differences between protein mutants (Gisselbrecht *et al*., 2013), and we anticipate it to be of great use in future bioinformatic studies.

Funding

This work was supported, in part, by the National Basic Research Program of China (2012CB316504), Shenzhen Special Fund for Strategic Emerging Industry grant ZD201111080127A and the National Science Foundation Grants DMS-1120368 and IIS-1017967.

Conflict of Interest: none declared.

References

Ackermann,M. and Strimmer,K. (2009) A general modular framework for gene set enrichment analysis. *BMC bioinformatics*, **10**, 47.

Aschard,H. *et al*. (2013) A nonparametric test to detect quantitative trait loci where the phenotypic distribution differs by genotypes. *Genet. Epidemiol.*, **37**, 323–333.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)*, **85**, 289–300.

Burke,D.T. *et al*. (2012) Dissection of complex adult traits in a mouse synthetic population. *Genome Res.*, **22**, 1549–1557.

Eddelbuettel,D. and François,R. (2011) Rcpp: seamless R and C++ Integration. *J. Stat. Softw.*, **40**, 118.

Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107–129.

Gisselbrecht,S.S. *et al*. (2011) Highly parallel assays of tissue-specific enhancers in whole *Drosophila* embryos. *Nat. Methods*, **10**, 774–780.

Goeman,J.J. and Bühlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Jiang,B. *et al*. (2014) Non-parametric *K*-sample tests via dynamic slicing. *J. Am. Stat. Assoc.*, in press.

Meyer,C.A. *et al*. (2011) BINOCh: Binding inference from nucleosome occupancy changes. *Bioinformatics*, **27**, 1867–1868.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA.*, **102**, 15545–15550.