Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Array Platform Modeling and Analysis (A)

CORM: An R Package Implementing the Clustering of Regression Models Method for Gene Clustering

Jiejun Shi and Li-Xuan Qin

Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

ABSTRACT: We report a new R package implementing the clustering of regression models (CORM) method for clustering genes using gene expression data and provide data examples illustrating each clustering function in the package. The CORM package is freely available at CRAN from http://cran.r-project.org.

KEYWORDS: clustering, gene expression, R package

SUPPLEMENT: Array Platform Modeling and Analysis (A)

CITATION: Shi and Qin. CORM: An R Package Implementing the Clustering of Regression Models Method for Gene Clustering. *Cancer Informatics* 2014:13(S4) 11–13 doi: 10.4137/CIN.S13967.

RECEIVED: April 15, 2014. **RESUBMITTED:** July 21, 2014. **ACCEPTED FOR PUBLICATION:** July 21, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Software Development

FUNDING: This work was supported by NIH grants CA151947 and CA008748 (LXQ and JS). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: qinl@mskcc.org

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties.

Introduction

Clustering genes by their expression profiles among samples and/or across time points is a useful approach to reducing data dimension and identifying co-expressed genes that may share common biological functions or regulatory networks.[1] There are three typical experimental designs, under which gene expression data are collected: (1) multiple samples at one time point, (2) one sample at multiple time points, and (3) multiple samples each at multiple time points. Data collected under each experimental design possess specific distributional characteristics to that design, which should be properly accounted for in data analysis.

For the analysis of gene clustering, we have previously developed a model-based clustering method, called the clustering of regression models (CORM) method, to accommodate data collected from various experimental designs while accounting for data distributional characteristics specific to each design.[2] CORM uses regression to model the expression of each gene and clusters genes that share similar regression coefficients to sample covariates. It applies an EM algorithm to iteratively assign genes to clusters and estimate the regression coefficients for each cluster. We have implemented CORM for the clustering of linear models (CLM) method and the clustering of linear mixed model (CLMM) method, with the former applied to cluster genes using cross-sectional data[2,3] and the latter time course data with or without replicates.[2,4]

In this paper, we report an R package that we recently developed implementing the CLM method and the CLMM method, and for each method, we provide two distinct data examples illustrating their uses.

CORM Package

The CORM package is available at R CRAN and it can be imported once installed with the R code below:

```
library('CORM')
```

Gene clustering for cross-sectional data. The CLM method is implemented by the function *fit.CLM*. Its usage is illustrated with two data examples in the package.

The first example is to cluster genes using microarray data derived from a set of breast cancer samples, including 38 invasive ductal carcinoma (IDC) and 21 invasive lobular carcinoma (ILC) samples.[5] Two indicator variables, one for IDC and another for ILC, were used as the covariates in the regression model for CLM. A set of 474 markers, which were selected based on their differential expression between the two breast cancer subtypes IDC versus ILC, are grouped into nine clusters with genes in each cluster sharing similar expression levels in both subtypes.[2]

The second example is to cluster candidate target genes for a microRNA named hsa-let-7f using their expression data derived from a set of normal and tumor tissue samples.[6] Three variables were used in the linear regression model for CLM: (1) let-7f expression, (2) indicator of disease status, and (3) interaction between let-7f and disease status. A set of 178 markers, which were selected based on their significant association with let-7f expression, are clustered to look for genes that share similar association with let-7f and hence may be similarly regulated by let-7f.[3]

Gene clustering for time course data. The CLMM method is implemented by two functions: *fit.CLMM* and *fit.CLMM.2*. The former can be used to cluster single-time course data or replicated time course data with replicate samples sharing the same time points. The latter can be used to cluster replicated time course with the samples belonging to two groups each having a different set of time points. They are each illustrated with a data example in the package.

A single-time course study examining yeast cell cycle is used to illustrate the use of *fit.CLMM*.[7] The fixed effects and random effects in the linear mixed effects model are both set to be the spline basis of time. A set of 256 cell cycle-dependent genes identified by Zhao et al.[8] were clustered into six groups. The expression profiles over time for genes in each group showed significant periodicity with similar peak time within each group and difference between groups.[2]

A replicated time course study assessing yeast cell cycle in two yeast samples of different genotypes is used to illustrate the use of *fit.CLMM.2*.[4] The two genotypes are wild-type yeast and single-mutant yeast with *YOX1* gene knocked out. The two samples were measured for gene expression at the same time points in the experiment; however, each sample incurred bad time points at different times due to technical issues, which were removed from the clustering analysis. In order to accommodate the different time points for the two samples, separate arguments for the fixed and random effects are allowed for the two samples in *fit.CLMM.2*. The same list of 256 cell cycle-dependent genes is partitioned into eight groups using the wild-type and mutant yeast data to look for genes whose expressions are similarly regulated by the mutation.[4]

Conclusion

CORM is an R package including a family of model-based clustering methods. It can be applied to cluster gene expression data collected under various types of experimental designs and forms an integrative analysis framework together with regression models typically used to detect differentially expressed or time-dependent genes. Table 1 lists the computing time spent for the aforementioned data examples as tested on a Linux server (Intel Xeon). In our experience, applying the CORM method, the EM algorithm typically converges in just a few iterations and is quite robust to the starting values when there are relatively well-separated clusters in the data. The user can also try multiple starting values by specifying the input parameter "n.start" in the CORM functions. We would recommend applying CORM to about 100–500 genes to partition them into reasonably sized clusters.

Author Contributions

Conceived and designed the experiments: LXQ. Analyzed the data: LXQ, JS. Wrote the first draft of the manuscript: JS. Contributed to the writing of the manuscript: LXQ, JS. Agree with manuscript results and conclusions: LXQ, JS. Jointly developed the structure and arguments for the paper: LXQ, JS. Made critical revisions and approved final version:

Table 1. Computing time by CORM functions versus by K-means clustering when applied to example datasets.

DATASET NAME	DATASET SIZE	FUNCTION NAME	COMPUTING TIME (sec)
Breast cancer[5]	354×59	fit.CLM	4.218
		kmeans	0.006
MicroRNA targets[6]	178×43	fit.CLM	120.363
		kmeans	0.004
Yeast cell cycle[7]	$256 \times 1 \times 16$	fit.CLMM	105.239
		kmeans	0.001
Wild type yeast cell cycle[4]	$256 \times 2 \times 24$	fit.CLMM.NA	644.845
		kmeans	0.004
Wild type and mutant yeast cell cycle[4]	$256 \times 2 \times 24 + 256 \times 2 \times 22$	fit.CLMM.NA.2	2558.402
		kmeans	0.006

LXQ, JS. Both authors reviewed and approved of the final manuscript.

REFERENCES

1. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*. 1998;95:14863–8.
2. Qin LX, Self SG. The clustering of regression models method with applications in gene expression data. *Biometrics*. 2006;62:526–33.
3. Qin LX. An integrative analysis of microRNA and mRNA expression – a case study. *Cancer Inform*. 2008;6:369–79.
4. Qin LX, Breeden L, Self SG. Finding gene clusters for a replicated time course study. *BMC Res Notes*. 2014;7:60.
5. Zhao H, Langerod A, Ji Y, et al. Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. *Mol Biol Cell*. 2004;15:2523–36.
6. Lu J, Getz G, Miska EA, et al. MicroRNA expression profiles classify human cancers. *Nature*. 2005;435:834–8.
7. Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycleregulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell*. 1998;9:3273–97.
8. Zhao LP, Prentice R, Breeden L. Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc Natl Acad Sci U S A*. 2001;98:5631–6.