Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

rScudo: an R package for classification of molecular profiles using rank-based signatures

Matteo Ciciani[1,*], Thomas Cantore[1], Mario Lauria[2,3]

[1]CIBIO – Centre for Integrative Biology, University of Trento, via Sommarive, 9, 38123 Povo (TN), Italy. [2]Department of Mathematics, University of Trento, via Sommarive, 14, 38123 Povo (TN), Italy. [3]The Microsoft Research - University of Trento Centre for Computational and Systems Biology (COSBI), Piazza Manifattura 1, 38068 Rovereto (TN), Italy.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: The classification of biological samples by means of their respective molecular profiles is a topic of great interest for its potential diagnostic, prognostic and investigational applications. rScudo is an R package for the classification of molecular profiles based on a radically new approach consisting in the analysis of the similarity of rank-based sample-specific signatures. The validity of rScudo unconventional approach has been validated through direct comparison with current methods in the international SBV IMPROVER Diagnostic Signature Challenge. Due to its novelty, there is ample room for conceptual improvements and for exploring additional applications. The rScudo package has been specifically designed to facilitate experimenting with the rank-based signature approach, to test its application to different types of molecular profiles, and to simplify direct comparison with existing methods.
Availability: The package is available as part of the Bioconductor suite at https://bioconductor.org/packages/rScudo
Contact: matteo.ciciani@studenti.unitn.it
Supplementary information: none.

Introduction

Given a set of biological samples, it is a problem of great interest how to develop a reliable method to classify the samples as belonging to a group or another based on a set of molecular profiles characterizing the samples; examples of molecular profiles are gene expression profiles, protein abundance, miRNA profiles. A sufficiently accurate classification method could be used for clinical applications such as a diagnostic test (i.e. sample belonging to a healthy or affected individual), patient stratification (sample belonging to group A of group B), or for investigational purposes (identification of dysregulated genes/pathway in individuals belonging to a certain group). rScudo is an R package for the classification of molecular profiles based on the analysis of their degree of similarity. Given a set of profiles, the method works by first extracting a rank-based signature from each profile. Signatures are then compared in a pairwise fashion to obtain a complete distance matrix, which is used to cluster profiles into groups; the grouping is used to reach a data-based classification ideally reflecting the desired phenotype-based one (such as control/disease status). The tool can be used either in supervised or unsupervised mode as detailed below. One significant advantage that derives from being rank-based is that the method does not require a preliminary normalization step as most other methods. rScudo implements the innovative approach to classification first demonstrated in the SCUDO online tool (Lauria, 2013; Lauria *et al.*, 2015), but adds several key features that expand its functionality and improve its useability:

- rScudo supports n-way sample classifications
- it offers a way to tune the values of the method parameters;
- it adds a method to automatically assign subjects to one of the user-defined groups, as opposed to just producing a similarity map of subjects requiring interpretation by the user;
- it simplifies the computation of classification performance metrics to facilitate direct comparisons with other methods;
- its highly optimized distance computation routine enables handling relatively large datasets;
- it allows customization of crucial parts of the algorithm (i.e. distance computation) to encourage experimenting with it.

Implementation and general features

A minimal script contains a call to the functions `scudoTrain`, `scudoNetwork` and `scudoPlot`. Function `scudoTrain` takes in input a set of profiles organized as a data frame with one feature (i.e. a gene, miRNA species or protein) per row and one profile per column; it returns a data structure (an object of class *scudoResults*) containing sample-specific gene signatures, consensus gene signatures for each group and the computed sample distance matrix. The signature length is constant for all samples and is specified by the user in terms of the required input arguments *nTop* and *nBottom*, corresponding respectively to the number of up-regulated and down-regulated features.

The returned distance matrix can then be used in a number of ways, the most immediate of which is to draw a sample similarity map. By treating the distance matrix as an adjacency matrix, the function `scudoNetwork` builds the map in the form of a graph with the profiles/samples represented as nodes, and only those distances falling below the N^{th} quantile represented as edges, where `N` is a required parameter of the function. The map can then be plotted within the R environment using the function `scudoPlot`. Alternatively, the function `scudoCytoscape` can be used to display the same map using the Cytoscape tool.

The expected result is a map showing samples spontaneously clustering together based on profile similarity. If this is indeed the case and the emerging clusters are clearly delineated, the classification can be considered successful and the class assignment of each sample trivially follows from the grouping. In less clear-cut cases, the map returned by `scudoNetwork` can be analyzed with the help of one of the many existing clustering/community identification methods (see for example the function `cluster_spinglass` from the package `igraph`) to algorithmically identify the groups.

In order to test the quality of the classification using an unrelated set of profiles (i.e. a testing set), the function `scudoTest` can be used. Such function takes as inputs both the new set of profiles and the object returned by `scudoTrain`, and performs a classification using only the features selected in the training step. The object returned by the call (an object of class *scudoResults*) can then be used to generate and plot a map of the testing set using the same functions `scudoNetwork` and `scudoPlot` as before.

The workflow just described shows how to perform an unsupervised classification of profiles using rScudo. An alternative workflow can be used for supervised classification, which works better for difficult to classify data sets, using the function `scudoClassify`. This function adds the samples to be classified to a reference map built with profiles for which the classification is known, and then uses a majority voting algorithm to infer the classification of each remaining sample based on that of its labelled neighbors. The function returns a list containing the predicted class for each sample in the test set plus a data frame of the classification scores used to generate the predictions.

All the above functions can handle two or more groups, thus making multi-group classification possible and extending the functionality of the online SCUDO tool. To encourage experimentation with variants of the algorithm, the `scudoTrain` and `scudoTest` have been parametrized with respect to the distance function used to build the distance matrix.

Another additional feature that has been implemented in the package is the optimization of the signature length, which was previously left to the user. Instead of a manual trial-and-error approach, the parameters *nTop* and *nBottom* can be automatically tuned using cross-validation combined to a grid search. For this purpose, the function `scudoModel` is used to generate a model object of the correct format for use as the *method* argument of the function `train` in the `caret` package. `Caret` is a well-known package that offers an extensible framework to perform parameters tuning using both grid search and cross-validation. The tuning of the parameters is not critical because we have shown that the method is robust to variations of the signature length (Lauria, 2013).

Conclusions

The validity of the rank-based signature approach has been validated by direct comparison with current methods in the international SBV IMPROVER Diagnostic Signature Challenge, organized by researchers from IBM Research and from Philip Morris International. Out of 52 competitors, the rank-based method placed second overall, reaching first place in one the four sub-challenges, namely the diagnosis of Multiple Sclerosis which interestingly appeared to be the most difficult of the four based on the empirical null distribution of scores (Tarca *et al.*, 2013; Norel *et al.*, 2013).

The method on which the package is based is agnostic about the nature of the profiles and the technology used to obtain them, and it has been shown to work on different types of high dimensional data. The rScudo package has been designed to facilitate experimenting with the rank-based approach to the profile classification problem, to test its application to different types of molecular profiles, and then to simplify direct comparison of results with existing methods. Representative examples of applications of the rScudo approach are the analysis of miRNA profiles for diagnostic purposes (Lauria, 2015), the integrative analysis of omics to decipher the role of APOE4 in Alzheimer's Disease (Caberlotto *et al.*, 2016), the study of gene expression profiles to gain insight into the molecular basis of nutritional interventions known to promote healthy aging (Lacroix *et al.*, 2015), the classification and study of protein profiles in Duchenne muscular dystrophy (Parolo *et al.*, 2018). Additional details on the use of rScudo and a detailed example can be found in the tutorial (vignette) available on its Bioconductor page (`https://bioconductor.org/packages/rScudo`).

Conflict of Interest: none declared.

References

Caberlotto,L. *et al.* (2016) Integration of transcriptomic and genomic data suggests candidate mechanisms for APOE4-mediated pathogenic action in Alzheimer's disease. *Sci. Rep.*, **6**, 32583.

Lacroix,S. *et al.* (2015) Systems biology approaches to study the molecular effects of caloric restriction and polyphenols on aging processes. *Genes Nutr.*, **10**, 58.

Lauria,M. (2015) Rank-based miRNA signatures for blood-based diagnosis of tuberculosis. *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, **2015-Novem**, 4462–4465.

Lauria,M. (2013) Rank-based transcriptional signatures: A novel approach to diagnostic biomarker definition and analysis. *Syst. Biomed.*, **1**, 35–46.

Lauria,M. *et al.* (2015) SCUDO: a tool for signature-based clustering of expression profiles. *Nucleic Acids Res.*, **43**, W188–92.

Norel,R. *et al.* (2013) sbv IMPROVER Diagnostic Signature Challenge. *Syst. Biomed.*, **1**, 208–216.

Parolo,S. *et al.* (2018) Combined use of protein biomarkers and network analysis unveils deregulated regulatory circuits in Duchenne muscular dystrophy. *PLoS One*, **13**.

Tarca,A.L. *et al.* (2013) Strengths and limitations of microarray-based phenotype prediction: Lessons learned from the IMPROVER Diagnostic Signature Challenge. *Bioinformatics*, **29**, 2892–2899.