BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 10 2013, pages 1348–1349
doi:10.1093/bioinformatics/btt136

Systems biology

Advance Access publication March 18, 2013

SRV: an open-source toolbox to accelerate the recovery of metabolic biomarkers and correlations from metabolic phenotyping datasets

Vincent Navratil[1,*,†], Clément Pontoizeau[1], Elise Billoir[2] and Benjamin J. Blaise[1,*]

[1]Centre de RMN à Très Hauts Champs, Institut des sciences analytiques, CNRS/ENS Lyon/UCB Lyon 1, Université de Lyon, 5 rue de la Doua, 69100 Villeurbanne, France and [2]Plateforme de Recherche de Rovaltain en Toxicologie Environnementale et Ecotoxicologie, 1 avenue de la Gare, BP 15173, 26958 Valence Cedex 9, France

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Supervised multivariate statistical analyses are often required to analyze the high-density spectral information in metabolic datasets acquired from complex mixtures in metabolic phenotyping studies. Here we present an implementation of the SRV—Statistical Recoupling of Variables—algorithm as an open-source Matlab and GNU Octave toolbox. SRV allows the identification of similarity between consecutive variables resulting from the high-resolution bucketing. Similar variables are gathered to restore the spectral dependency within the datasets and identify metabolic NMR signals. The correlation and significance of these new NMR variables for a given effect under study can then be measured and represented on a loading plot to allow a visual and efficient identification of candidate biomarkers. Further on, correlations between these candidate biomarkers can be visualized on a two-dimensional pseudospectrum, representing a correlation map, helping to understand the modifications of the underlying metabolic network.

Availability: SRV toolbox is encoded in MATLAB R2008A (Mathworks, Natick, MA) and in GNU Octave. It is available free of charge at http://www.prabi.fr/redmine/projects/srv/repository with a tutorial.

Contact: benjamin.blaise@chu-lyon.fr or vincent.navratil@univ-lyon1.fr

Received and revised on February 20, 2013; accepted on March 12, 2013

1 INTRODUCTION

Given the complexity of metabolic samples used in metabonomics to understand the metabolic response of an organism to pathophysiological stimuli (Nicholson *et al.*, 1999), the development of efficient approaches to fully analyze datasets is a growing field (Lavine and Workman, 2010). An NMR spectrum typically contains a few hundreds of resonances over a 10-ppm-wide chemical shift range. A high-resolution bucketing, using 0.001-ppm-wide buckets, leads to an efficient sampling of the signal, matching the high resolution of the NMR spectra, and results in a few thousands of variables to handle for further statistical analyses. Multivariate statistical analyses, such as orthogonal partial least squares (O-PLS) regressions (Trygg and Wold, 2002), are thus efficient tools to explore these high-density information datasets and combine collective modifications of metabolite concentrations to allow a discrimination of the samples with respect to the effect under study.

However, the interpretation of the latent variables, which are the results of multivariate statistical analyses, is far from trivial. The mainly used approach is to combine the representation of the latent variable with the Pearson correlation coefficients of each variable with the information matrix, encoding the different classes (Cloarec *et al.*, 2005). The difficulty with this approach is that multiple buckets represent a single NMR peak. It thus shows different levels of correlation, despite the fact that it represents a single metabolic signal. A second difficulty is the definition of a correlation threshold above which a signal can be considered as valuable to designate a candidate biomarker.

Statistical Recoupling of Variables (SRV) is an algorithm designed to overthrow these main difficulties (Blaise *et al.*, 2009). The conducting idea is to restore the spectral dependency that was lost by high-resolution bucketing. The statistical relationships between consecutive variables allow aggregating them into clusters following the highest direction of covariance/correlation ratio, thus defining NMR peaks. Neighbouring clusters can then be merged into superclusters to recover NMR multiplets. These superclusters correspond to NMR variables of interest. SRV thus acts as an automated variable-size bucketing procedure coupled with an efficient noise-removing filter.

We then use a significance-testing filter using multiple hypothesis testing corrections. The Benjamini–Yekutieli measurement of the false discovery rate seems adapted to NMR-based metabonomics (Benjamini and Yekutieli, 2001), but other less strict corrections could be considered. A typical threshold of 0.05 can then be defined, and a simple identification of statistically significant signals allows the recovery of candidate biomarkers.

Based on the statistical total correlation spectroscopy (Cloarec *et al.*, 2005), it is possible to establish a two-dimensional (2D) pseudospectrum defined as a correlation map between the superclusters. These correlations can eventually be interpreted on the global metabolic network to extract the perturbed metabolic network associated with a major or minor perturbation (Blaise *et al.*, 2010, 2011). Here we present an open-source implementation of the SRV algorithm as MATLAB/GNU Octave functions leading to the visualization of the latent variables after O-PLS

*To whom correspondence should be addressed.
†Present address: Pôle Rhône Alpes de Bioinformatique, Université Lyon 1, Bâtiment Gregor Mendel, 16 rue Raphaël Dubois, 69100 Villeurbanne, France.

analysis for the discrimination between two groups, with the correlation and significance testing representation, and the 2D pseudospectrum allowing the identification of coordinated metabolic variations.

2 DESCRIPTION OF THE SRV ALGORITHM AND IMPLEMENTATION

The SRV algorithm is divided into five steps, as schematically described in the following text (Blaise *et al.*, 2009). The first three steps correspond to the statistical mining of metabolite biomarker signals (i.e. the SRV clusters) from a set of NMR spectra.

Step 1: Definition of a spectral dependency landscape (L) as the covariance/correlation ratio between neighbouring variables along the chemical shift axis:

$$L(i) = \frac{\text{covariance}}{\text{correlation}}(i, i+1) = \sqrt{\text{variance(i)} \times \text{variance(i+1)}}$$
$$= \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N}(i-\bar{i})^2\right) \times \left(\frac{1}{N}\sum_{i=1}^{N}((i+1)-(\overline{i+1}))^2\right)}$$

Step 2: Identification of spectral SRV clusters.

(i) The first variable of the dataset starts the first cluster.
(ii) The spectral dependency landscape is scanned to identify local minima of covariance/correlation ratio that represent the borders between two clusters.
(iii) Clusters representing NMR signals are defined by a minimum number of variables, which depends on the resolution of the NMR spectra.

Step 3: Identification of NMR variables.

(iv) Superclusters are based on the aggregation of clusters depending on their correlation with their neighbouring clusters.
(v) The intensity of the supercluster is the mean of the intensities of the NMR signal in the buckets assigned to the supercluster.

Step 4: Evaluation of *P*-values and multiple dependent tests correction using the Benjamini–Yekutieli false discovery rate (Benjamini and Yekutieli, 2001). An adjusted *P*-value threshold is estimated by the identification of the highest rank verifying the equation below, where N is the number of variables used in the model. We then can reject all null hypotheses corresponding to rank 1 to k.

$$k = \max\left(i = 1 : N, p_i < \frac{i \times 0.05}{N \times \sum_{i=1}^{N}\frac{1}{i}}\right)$$

The command line is executed as follow: [Data, Xclusterf, Ibegin, Iend, number of clusters] = SRV (X matrix, Y matrix, typical singlet peak base width, bucketing resolution, correlation threshold, significance threshold, ppm, number of factors).

We commonly use the analysis of variance for the evaluation of *P*-values and the Benjamini–Yekutieli correction for the measurement of the false discovery rate in our NMR metabolic phenotyping datasets. However, other evaluation procedures can be used on SRV clusters based on the properties of the variables under study. SRV clusters are available in the output Xclusterf of the SRV function. Data is a four-row table containing the ppm line, the loading value, the correlation value and the significance of each initial NMR variable. Xclusterf is a matrix containing the intensity of the signal in each cluster for the different spectra of the dataset. Ibegin and Iend are tables containing the limits of each cluster, and S is a table allowing the identification of the initial NMR signal contained in the SRV clusters and the amount of signal lost. X matrix is the dataset matrix with spectra in row. Columns of zeros must represent the excluded residual water signal area. SRV is not able to deal with multiple exclusion areas (for instance, NMR buffer signals). In such cases, additional exclusion areas should simply be removed from the dataset before the use of SRV. One can eventually choose to reintroduce these areas in the output matrix of SRV. For the corresponding chemical shifts, O-PLS coefficients and correlation values should be put to 0 and *P*-values to 1. Y matrix is the column vector encoding the membership of each sample to the groups under study. ppm is a row vector containing the ppm value of each bucket. Number of factors is the total number of components for the O-PLS analysis, including the orthogonal ones. The main output of the function is the loading plot of the O-PLS analysis (Cloarec *et al.*, 20054) with a colour code of correlation. Variables that are not statistically significant are coloured in grey.

Step 5: Computation and visualization of the SRV 2D pseudospectrum. Finally, after a deflation of the data matrix by an O-PLS analysis based on the class information matrix (Trygg and Wold, 2002), we compute the autocorrelation matrix between SRV clusters identified within the dataset (Cloarec *et al.*, 2005).

[ORStocsy, Correlation table] = orstocsy (Xclusterf, Y matrix, number of factors, correlation threshold, Ibegin, Iend, ppm).

ORStocsy is a $N \times N$ matrix containing the correlation between SRV clusters (Blaise *et al.*, 2010, 2011). Correlation table is a seven-row table containing the identification number of the correlated clusters, the ppm starting and ending values of the correlated clusters and the level of correlation. The other parameters are identical to those described previously.

Funding: The French government.

Conflict of Interest: none declared.

REFERENCES

Blaise,B.J. *et al.* (2009) Statistical recoupling prior to significance testing in nuclear magnetic resonance based metabonomics. *Anal. Chem.*, **81**, 6242–6251.
Blaise,B.J. *et al.* (2010) Two-dimensional statistical recoupling for the identification of perturbed metabolic networks from NMR spectroscopy. *J. Proteome Res.*, **9**, 4513–4520.
Blaise,B.J. *et al.* (2011) Orthogonal Filtered Recoupled-STOCSY to extract metabolic networks associated with minor perturbations from NMR spectroscopy. *J. Proteome Res.*, **10**, 4342–4348.
Benjamini,B.Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.
Cloarec,O. *et al.* (2005) Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Anal. Chem.*, **77**, 1282–1289.
Lavine,B. and Workman,J. (2010) Chemometrics. *Anal. Chem.*, **82**, 4699–4711.
Nicholson,J. *et al.* (1999) "Metabonomics": understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, **29**, 1181–1189.
Trygg,J. and Wold,S. (2002) Orthogonal projections to latent structures (O-PLS). *J. Chemom.*, **16**, 119–128.