BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 20 2006, pages 2562–2564
doi:10.1093/bioinformatics/btl428

Systems biology

HiRes—a tool for comprehensive assessment and interpretation of metabolomic data

Qi Zhao[1], Radka Stoyanova[2], Shuyan Du[3], Paul Sajda[3] and Truman R. Brown[1,3,*]

[1]Hatch Center for MR Research, Department of Radiology, Columbia University, New York, NY 10032 USA, [2]Department of Radiation Oncology, Fox Chase Cancer Center, Philadelphia, PA 19111 USA and [3]Department of Biomedical Engineering, Columbia University, New York, NY 10027 USA

Received on April 10, 2006; revised on June 14, 2006; accepted on August 1, 2006
Advance Access publication August 7, 2006
Associate Editor: Nikolaus Rajewsky

ABSTRACT

Summary: The increasing role of metabolomics in system biology is driving the development of tools for comprehensive analysis of high-resolution NMR spectral datasets. This task is quite challenging since unlike the datasets resulting from other 'omics', a substantial preprocessing of the data is needed to allow successful identification of spectral patterns associated with relevant biological variability. *HiRes* is a unique stand-alone software tool that combines standard NMR spectral processing functionalities with techniques for multi-spectral dataset analysis, such as principal component analysis and non-negative matrix factorization. In addition, *HiRes* contains extensive abilities for data cleansing, such as baseline correction, solvent peak suppression, removal of frequency shifts owing to experimental conditions as well as auxiliary information management. Integration of these components together with multivariate analytical procedures makes *HiRes* very capable of addressing the challenges for assessment and interpretation of large metabolomic datasets, greatly simplifying this otherwise lengthy and difficult process and assuring optimal information retrieval.

Availability: *HiRes* is freely available for research purposes at http://hatch.cpmc.columbia.edu/highresmrs.html

Contact: qz2106@columbia.edu

1 INTRODUCTION

Metabolomics (or metabonomics) has been labeled one of the new 'omics', joining genomics, transcriptomics, and proteomics as a science employed towards the understanding of global systems biology (Nicholson and Wilson, 2003). Metabolomics is defined here as the quantitative measurement of the dynamic, multiparametic metabolite response of living systems to pathophysiological stimuli or genetic modification. The analysis of the metabolome is particularly challenging owing to the diverse chemical nature of metabolites and thus the wide array of analytical methodology employed in their detection. NMR spectroscopy is one of the mainstays in this array and provides a wide range of information for metabolic characterization of biological samples. However, the quantity and complexity of spectroscopic data obtained in metabolomic studies have made data interpretation very difficult. The challenges stem not only from the need to apply standard processing procedures to large sets of spectra and connect them with auxiliary information about the samples, but also the frequent necessity to invoke pattern recognition methods to ensure optimal information retrieval (Nicholson *et al*., 1999). Most of the existing software deals only with a few of the aspects of this process and often acts on a single spectrum.

Here we present a software tool, High Resolution Spectroscopy (*HiRes*), which addresses these challenges and provides comprehensive analysis of metabolomic datasets. It combines standard spectral processing routines, data correction functions, techniques for reducing information complexity of multi-spectral dataset such as principal component analysis (PCA) (Stoyanova and Brown 2001) and constrained non-negative matrix factorization (cNMF) (Sajda 2004). PCA is an invaluable aid in the exploration of large datasets, allowing representation of complex data in lower dimensional space, defined by the principal components (PCs). cNMF is a pattern recognition algorithm that extracts a set of spectral features with direct physical interpretation, which can be used for identifying meaningful biochemical effects. It is based on non-negative matrix factorization (Lee and Seung, 1996, 1999), where the recovered spectral patterns define a subspace that envelops the observed spectral data with minimal error. In cNMF, there is a further constraint that the underlying spectral patterns and their corresponding strengths are forced to be positive, thus physically realizable and interpretable. The variation of the strengths of these spectral features reflects potential metabolic changes across the dataset, corresponding to disease status or treatment end points.

2 FUNCTIONALITY

HiRes is a software developed using C++ on Windows platform. Its functionality is illustrated on a dataset, the analysis of which we have reported previously (Stoyanova *et al*., 2004a,b). With *HiRes* the time needed to reproduce these results was less than 10 minutes. The analyzed dataset consists of 57 ^{1}H NMR spectra from rat urine, obtained during dose- and time-related experiments with a well-known liver toxin—hydrazine. The rats in the experiment are divided into four groups, one serving as a control and the remaining three treated with three different doses of hydrazine. Urine is drawn before hydrazine admission (0 h) and every 8 h post-dose. Detailed description of the experiment and data acquisition can be found elsewhere (Nicholls *et al*., 2001).

*To whom correspondence should be addressed.

 © The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from http://bioinformatics.oxfordjournals.org/ at NERL on December 14, 2016

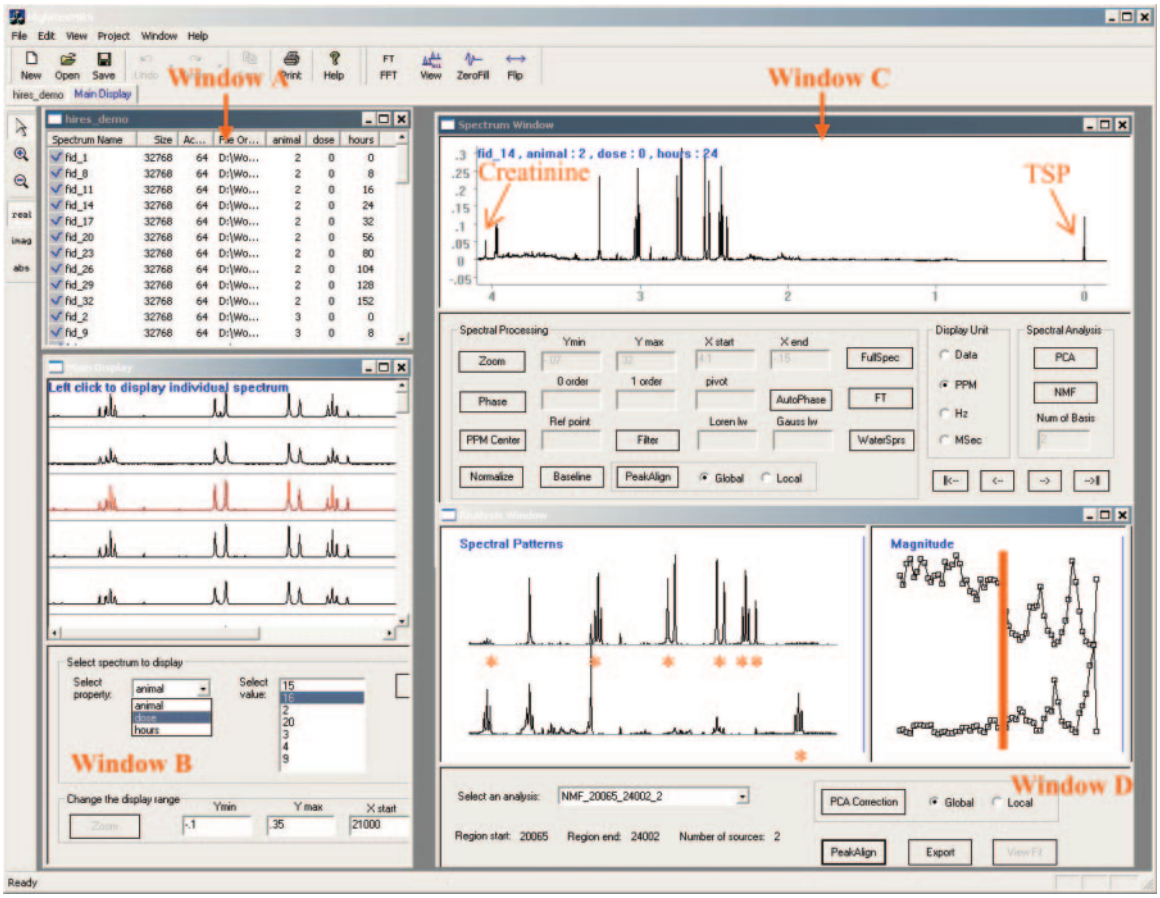

Fig. 1. Screenshot of *HiRes* based on the analysis of ^{1}H NMR spectra from rat urine, illustrating the functionalities of the program from four aspects: data importation and organization (window A), multiple spectra display (window B), spectral processing utilities (window C) and pattern discovery (window D).

Data importation and organization

HiRes imports a series of NMR spectra in a single step, together with any additional information associated with the data from a separate tab-delimited text file. All the spectra and associated information are organized in a project and listed in a table that can be easily sorted. Datasets from multiple studies can be imported and processed separately, and then combined into one project for cross-study analysis. A screenshot of *HiRes* is presented in Figure 1. In the case of the hydrazine dataset, the spectra are associated with and also sorted by measurement time, animal ID and dose (window A). *HiRes* allows automatic importation of spectra acquired from Bruker (Bruker Biospin, Karlsruhe, Germany) and Varian (Varian Inc., Palo Alto, USA) spectrometers. Future versions would also include support for JEOL (JEOL Ltd, Tokyo, Japan) spectrometers, as well as processed spectra from variety of existing software. In addition, spectral parameters can be manually specified for any other data format. Multiple spectra are grouped together and labeled for easy browsing (window B) and can be selectively displayed by specifying property filters. Spectra can be exported at any stage for saving intermediate results.

Spectral processing

The spectral processing parameters are determined by individual examination of the spectra (window C). *HiRes* provides various interactive data processing procedures, which can be either applied to all the spectra with the same set of parameters or to each spectrum with different parameters. All the processing steps are tracked in a history list and can be undone or redone. These functions include: Fast Fourier Transform (FFT), constant and linear phase correction (manually or automatically); temporal filtering (Lorentzian or Gaussian); baseline correction using convolution difference or linear fitting; zero filling; water suppression; normalization using single reference peak, subset or entire spectrum. In the case of the hydrazine dataset the spectra are normalized by the integral of the creatinine [4.0–4.1 p.p.m.].

Frequency shifts correction

Often the performance of pattern recognition is impeded by experimental and instrument-induced variations which in general obscure the process of pattern discovery. Frequency shifts are the dominant source of such unwanted variations in spectral datasets. *HiRes* provides global shifting of the entire spectrum based on the alignment of an individual reference peak. The data in the hydrazine dataset are initially aligned using the TSP resonance [0 p.p.m.]. However, some individual peaks in the dataset are still misaligned, presumably owing to small variations in solution conditions in the urine samples. These variations are identified in *HiRes* by examining the presence of peak-derivative shapes in the second and higher order

Downloaded from http://bioinformatics.oxfordjournals.org/ at NERL on December 14, 2016

PCs (Stoyanova, 2004a). The corresponding peak-regions are then selected and the local frequency shifts adjusted within these spectral regions. This eliminates the need of 'binning' spectra to only a few hundred points, which has been a standard way to deal with local frequency variations until now.

Spectral pattern recognition

After the spectra are preprocessed the underlying spectral patterns can be explored using PCA and cNMF. In the left panel in window D the two spectral patterns recovered via cNMF in the region [2.1–3.5 p.p.m.] are displayed. Note that these are physically realizable spectra and can be directly interpreted in terms of a common varying set of metabolites. The stars denote the resonance regions where peaks were adjusted following the alignment procedure. At the right panel in window D the corresponding magnitudes of the spectral patterns are displayed. Each point of the magnitude plot represents the strength of the corresponding spectral pattern in an individual data spectrum. The display order of these magnitudes will be updated automatically if the spectra are re-sorted by different property. In the example, the magnitudes indicate that the first pattern is high for the control spectra (those to the left of the orange bar) and decreases with the administration of the toxin. In contrast, the second pattern is low for the controls and changes with hydrazine dose and time following administration. Thus, the first pattern is associated with the urine unaffected by the hydrazine, and the second is reflecting the changes owing to the hydrazine administration. The spectral patterns can be directly related to the biochemical changes owing to the hydrazine effects: the 'normal' pattern (upper row) contains principally the peaks from citrate, succinate, 2-oxoglutarate and trimethylamine-*N*-oxide, while the lower pattern consists mainly of 2-aminoadipic acid, β-alanine, creatine and taurine. More complex patterns can be recovered if three or four solutions are sought in cNMF.

By right clicking on any point in the magnitude plot, the user has a choice to view or remove the associated individual spectrum, or apply any further correction that is necessary. Each time a spectrum is removed or modified, the analysis results are automatically updated. This direct interaction between the preprocessing and analysis stages greatly facilitates the pattern discovery process and identification of potential biomarkers.

3 SUMMARY

To the best of our knowledge, *HiRes* is the first software tool able to integrate the necessary preprocessing and analysis steps for large NMR spectral datasets. *HiRes* couples rigorous data pre-processing, artifact removal and identification of metabolic patterns via PCA. The addition of the recently developed cNMF analysis procedure, which identifies biochemically meaningful and physically interpretable spectral patterns and mixtures, greatly aids the data interpretation. These patterns and mixtures provide significantly more insights into the underlying metabolic behavior of the system under study. *HiRes* is easy to use, well-documented and we believe that data analysts and NMR-users will find utilizing its functionality very intuitive.

ACKNOWLEDGEMENTS

Metabolic roadmap initiative of National Institutes of Health DK070301.

Conflict of Interest: none declared.

REFERENCES

Lee,D.D. and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.

Lee,D.D. and Seung,H.S. (1996) Unsupervised learning by convex and conic coding. *NIPS*, 515–521.

Nicholls,A.W. *et al*. (2001) Metabonomic investigations into Hydrazine toxicity in the rat. *Chem. Res. Toxicol*, **14**, 975–987.

Nicholson,J.K. and Wilson,I.D. (2003) Understanding 'Global' systems biology: Metabonomics and the continuum of metabolism. *Nat. Rev. Drug. Discov*., **2**, 668–676.

Nicholson,J.K. *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.

Sajda,P. *et al*. (2004) Non-negative matrix factorization for rapid recovery of constituent spectra in magnetic resonance spectroscopy imaging of the brain. *IEEE Trans. Med. Imaging*, **23**, 1453–1465.

Stoyanova,R. and Brown,T.R. (2001) NMR spectral quantitation by principal component analysis. *NMR Biomed*., **14**, 271–277.

Stoyanova,R. *et al*. (2004a) Automatic alignment of individual peaks in large high-resolution spectral data sets. *J. Magn. Res*., **170**, 329–335.

Stoyanova,R. *et al*. (2004b) Sample classification based on Bayesian spectral decomposition of metabonomic NMR data sets. *Anal. Chem*., **76**, 3666–3674.

Downloaded from http://bioinformatics.oxfordjournals.org/ at NERL on December 14, 2016