BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 6 2011, pages 885–886
doi:10.1093/bioinformatics/btr013

Data and text mining

Advance Access publication January 6, 2011

The CCPN Metabolomics Project: a fast protocol for metabolite identification by 2D-NMR

Francesca Chignola[1], Silvia Mari[1,*], Tim J. Stevens[2,*], Rasmus H. Fogh[2], Valeria Mannella[1], Wayne Boucher[2] and Giovanna Musco[1]

[1]Biomolecular NMR Laboratory, Center of Genomics, BioInformatics and BioStatistics, Dulbecco Telethon Institute c/o San Raffaele Scientific Institute – DIBIT, Via Olgettina 58, 20132 Milan, Italy and [2]Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Old Addenbrooke's Site, Cambridge CB2 1GA, UK

Associate Editor: Alex Bateman

ABSTRACT

Summary: We present here the freely available Metabolomics Project resource specifically designed to work under the CcpNmr Analysis program produced by CCPN (Collaborative Computing Project for NMR) (Vranken *et al.*, 2005, The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins*, **59**, 687–696). The project consists of a database of assigned 1D and 2D spectra of many common metabolites. The project aims to help the user to analyze and assign 1D and 2D NMR spectra of unknown metabolite mixtures. Spectra of unknown mixtures can be easily superimposed and compared with the database spectra, thus facilitating their assignment and identification.

Availability: The CCPN Metabolomics Project, together with an annotated example dataset, is freely available via: http://www.ccpn.ac.uk/metabolomics/.

Contact: mari.silvia@hsr.it (experiments & protocol); tjs23@cam.ac.uk (software).

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on November 13, 2010; revised and accepted on January 4, 2011

1 INTRODUCTION

Metabolomics is a newly emerging field of 'omics' research aiming to give an analytical description of the metabolites in complex biological samples and to describe the global, dynamic metabolic response of living systems to biological stimuli or genetic manipulation (Nicholson *et al.*, 2008).

NMR spectroscopy is making significant contributions in this burgeoning area: it is successfully being used to identify biomarkers in biofluids for various diseases, to analyze drug toxicity and to determine drugs' *in vivo* efficacy and selectivity (Bernini *et al.*, 2009; Power 2009). Its potential ability to handle complex mixtures of metabolites makes it a primary tool in both identification and quantification of the multitude of different molecules that constitute unprocessed biological mixtures.

1D ^{1}H NMR spectra of biological mixtures are characterized by extensive peak overlap or spectral congestion (Xia *et al.*, 2008), therefore acquisition of 2D NMR spectra (e.g. ^{1}H-^{13}C HSQC, HMBC and ^{1}H-^{1}H TOCSY) is a prerequisite for compound identification. However, the paucity of metabolomics software with dedicated 2D metabolite spectrum resources and coupled analysis capabilities often makes assignment of metabolite resonances a challenging task.

We present here the Metabolomics Project dataset based on CcpNmr Analysis software, consisting of a series of 1D and 2D NMR spectra of 76 standard compounds whose resonances (^{1}H and ^{13}C) have been fully assigned and annotated. This repository can be efficiently applied to the rapid identification of the metabolic components of unknown mixtures (Supplementary Tables S1–S3).

2 PROJECT OVERVIEW

The CCPN Metabolomics Project, available for the CcpNmr software suite (available from the Collaborative Computing Project for NMR, www.ccpn.ac.uk), consists of assigned 1D and 2D spectra of standard compounds. The project will be periodically updated with new metabolite entries. All spectra are fully assigned (Supplementary Fig. S1), according to public databases, including the Biological Magnetic Resonance Data Bank (Seavey *et al.*, 1991) and the Human Metabolome Database (Wishart *et al.*, 2007). The CCPN Metabolomics Project is opened by version 2.1 and above of CcpNmr Analysis software (Vranken *et al.*, 2005) (freely downloadable at http://www.ccpn.ac.uk/download/). Here, we present an overview of the CCPN Metabolomics Project, whose detailed description is available as a Tutorial in Supplementary Material.

2.1 Creation of the database

The metabolites in the database have been selected from the most common metabolites present in biofluids. For each metabolite, a description of the molecular system has been created, either downloading it from the CCPN repository (http://www.ebi.ac.uk/pdbe/docs/NMR/chemCompXml/main.html) at the EBI/PDBe or importing it as .mol file using CcpNmr FormatConverter feature (Chapter 10 of Tutorial in Supplementary Material). 1D-^{1}H spectra and 2D ^{1}H-^{13}C HSQCs downloaded from public databases satisfy the standard conditions for metabolomic purposes (pH 7, $T = 298$K) (Beckonert *et al.*, 2007). Spectra have been processed using Topspin 2.0 (Bruker BioSpin GmbH, Rheinstetten, Germany) and referenced using the standard methyl groups' peak of DSS (2,2-Dimethyl-2-silapentane-5-sulfonic acid). Importantly, ^{1}H-^{1}H TOCSY data, not available on public

*To whom correspondence should be addressed.

repositories, have been generated as simulated peak lists only, based on chemical structures and on the chemical shift values from the corresponding ^{1}H-^{13}C HSQC NMR spectra. In essence, the predicted TOCSY peaks are located at the ^{1}H-^{1}H chemical shift intersections, within spectrum bounds, for hydrogen atoms that are connected with one another as part of a theoretical 3J-linked network. Peak intensity information is not simulated.

2.2 Getting started

In order to start, four important steps should be performed: (i) opening the spectra of the unknown mixture (from now on referred to as 'query spectra'). Most common NMR spectrum data formats (such as Varian, Bruker, NMRpipe) can be read directly; (ii) referencing the query spectra (acquired at 298K, pH 7) to DSS; (iii) defining associated molecular systems. It is possible to select all molecular systems present in the project library or just those metabolites the users considers relevant for their specific project (Fig. 6 of Tutorial in Supplementary Material); (iv) creating a query peak list using an automatic or manual peak picking procedure on the spectra, prior to proceeding with the identification of metabolites.

2.3 Identification of metabolites in the query spectrum

Peak assignment can be performed easily in a supervised procedure by transferring the standard compounds' peak lists to the query peak list using the CcpNmr Analysis 'Copy Assignments' utility. A visual inspection is recommended to verify the correctness of assignment transfer (Fig. 12 of Tutorial in Supplementary Material). Alternatively, it is possible to assign an unknown peak by taking advantage of the CcpNmr Analysis 'Assignment Panel', which suggests assignment possibilities based on chemical shifts and one-bond connectivity. For the analysis of a 2D spectrum, it is possible to restrict the suggested assignments to intra-molecular contributions only (Fig. 13 of Tutorial in Supplementary Material). In order to confirm the assignment, the spectra of the corresponding standard metabolite(s) can be overlaid onto the query spectrum.

The procedures described here are feasible for all the spectrum windows, including the 1D ^{1}H and the 2D ^{1}H-^{1}H TOCSY. Of note is the possibility to include mono-dimensional spectra in the CCPN Metabolomics Project, which allows the detection of low concentration (up to 10 μM) metabolites.

2.4 Quality check

The resonance identification performance can be checked using the CcpNmr utility 'Quality Reports' (Chapter 9 of Tutorial in Supplementary Material) highlighting potential inconsistencies, errors and chemical shifts with unusual values or high standard deviation. Resonances with potential problems can be readily linked to their associated peaks. Finally, the 'Peak Lists' table allows a useful overview of the completeness of the assignment, and can be easily exported.

2.5 Customizing the library

The metabolite library can be personalized by adding new molecules and their corresponding spectra to the CCPN Metabolomics Project. A new standard compound molecule can be introduced by downloading it from the 'small compound' database available in CcpNmr Analysis. Alternatively, it can be imported as a .mol, .mol2 or .pdb file through the CcpNmr FormatConverter program (Chapter 10 of Tutorial in Supplementary Material).

2.6 Testing and validation

The CCPN Metabolomics Project has been tested on 2D ^{1}H-^{13}C HSQC spectra acquired on a cell culture supernatant collected at two different time points and on a mouse urine sample (Supplementary Tables S1–S3). Metabolite recognition under the CCPN Metabolomics Project performs as well as the identification obtained using the open-access software MetaboMiner (Xia *et al.*, 2008). Also, by using the CCPN Metabolomics Project users can easily handle real and fully assigned NMR spectra, both 2D and 1D, in a general NMR software suite, with all the extra analytical tools and the active support, for both developers and users, that CCPN provides.

3 FUTURE PERSPECTIVES

The CCPN Metabolomics Project is undergoing continuous development in order to improve its application in NMR metabolomics studies. Updated metabolomics projects with new reference spectra and reference assignments will be available and labeled at http://www.ccpn.ac.uk/metabolomics/. Availability of standard compounds will be increased and made available by developing a system to automatically update and select metabolite data from a database. Moreover, the project will be adapted to also manage 2D ^{1}H J-resolved NMR spectra, reducing overlapped signals and assignment ambiguity.

ACKNOWLEDGEMENTS

We thank Dr Boletta and Dr Pluchino for providing mouse urine samples and the cell culture media sample, respectively.

Funding: Fondazione Telethon and Fondazione Cariplo (to G.M.); CCPN project is supported by the Biotechnology and Biological Sciences Research Council.

Conflict of Interest: none declared.

REFERENCES

Beckonert,O. *et al*. (2007) Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat. Protoc.*, **2**, 2692.

Bernini,P. *et al*. (2009) Individual human phenotypes in metabolic space and time. *J. Proteome Res.*, **8**, 4264–4271.

Nicholson,J.K. and Lindon,J.C. (2008) Systems biology: metabonomics. *Nature*, **455**, 1054–1056.

Powers,R. (2009) NMR metabolomics and drug discovery. *Magn. Reson. Chem.*, **47**, S2–S11

Seavey,B.R. *et al*. (1991) A relational database for sequence-specific protein NMR data. *J. Biomol. NMR*, **1**, 217–236.

Vranken,W.F. *et al*. (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins*, **59**, 687–696.

Wishart,D.S. *et al*. (2007) HMDB: the human metabolome database. *Nucleic Acids Res.*, **35**, D521–D526.

Xia,J. *et al*. (2008) MetaboMiner–semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC Bioinformatics*, **9**, 507.