Pacific Symposium on Biocomputing 12:157-168(2007)

NEW BIOINFORMATICS RESOURCES FOR METABOLOMICS

JOHN L. MARKLEY, MARK E. ANDERSON, QIU CUI, HAMID R. EGHBALNIA,[*]
IAN A. LEWIS, ADRIAN D. HEGEMAN, JING LI, CHRISTOPHER F. SCHULTE,
MICHAEL R. SUSSMAN, WILLIAM M. WESTLER, ELDON L. ULRICH,
ZSOLT ZOLNAI

Department of Biochemistry, University of Wisconsin-Madison, 433 Babcock Drive, Madison, Wisconsin 53706, USA

We recently developed two databases and a laboratory information system as resources for the metabolomics community. These tools are freely available and are intended to ease data analysis in both MS and NMR based metabolomics studies. The first database is a metabolomics extension to the BioMagResBank (BMRB, http://www.bmrb.wisc.edu), which currently contains experimental spectral data on over 270 pure compounds. Each small molecule entry consists of five or six one- and two-dimensional NMR data sets, along with information about the source of the compound, solution conditions, data collection protocol and the NMR pulse sequences. Users have free access to peak lists, spectra, and original time-domain data. The BMRB database can be queried by name, monoisotopic mass and chemical shift. We are currently developing a deposition tool that will enable people in the community to add their own data to this resource. Our second database, the Madison Metabolomics Consortium Database (MMCD, available from http://mmcd.nmrfam.wisc.edu/), is a hub for information on over 10,000 metabolites. These data were collected from a variety of sites with an emphasis on metabolites found in *Arabidopsis*. The MMC database supports extensive search functions and allows users to make bulk queries using experimental MS and/or NMR data. In addition to these databases, we have developed a new module for the Sesame laboratory information management system (http://www.sesame.wisc.edu) that captures all of the experimental protocols, background information, and experimental data associated with metabolomics samples. Sesame was designed to help coordinate research efforts in laboratories with high sample throughput and multiple investigators and to track all of the actions that have taken place in a particular study.

1. Introduction

The metabolome can be defined as the complete inventory of small molecules present in an organism. Its composition depends on the biological fluid or tissue studied and the state of the organism (health, disease, environmental challenge, etc). Metabolomics is the study of the metabolome, usually as a high-throughput activity with the goal of discovering correlations between metabolite levels and the state of the organism. Metabolomics holds a place in systems biology

[*] Also Department of Mathematics, University of Wisconsin-Madison.

alongside genomics, transcriptomics, and proteomics as an approach to modeling and understanding reaction networks in cells [1–4].

Mass spectrometry (MS) and nuclear magnetic resonance (NMR) are the analytical techniques used in the majority of metabolomics studies [5, 6]. Although MS and NMR suffer from some well documented technical limitations [7], both of these tools are of clear utility to modern metabolomics [8]. MS is now capable of detecting molecules at concentrations as low as 10^{-18} molar, and high-field nuclear magnetic resonance (NMR) can efficiently differentiate between molecules that are as similar in structure as glucose and galactose.

Despite the availability of these impressive analytical tools, determining the molecular composition of complex mixtures is one of the most difficult tasks in metabolomics. One reason for this difficulty is a lack of publicly available tools for comparing experimental data with the existing literature on the masses and chemical shifts of common metabolites. We recently developed two databases of biologically relevant small molecules as practical tools for MS- and NMR-based research.

The first of these databases is a metabolomics extension to the existing Biological Magnetic Resonance Data Bank (BioMagResBank, BMRB). The BMRB database contains experimental NMR data from over 270 pure compounds collected under standardized conditions. The peak lists, processed spectra, and raw time-domain data are freely available at http://www.bmrb.wisc.edu. Although the initial data were collected by the Madison Metabolomics Consortium (MMC), several groups in the metabolomics community have expressed interest in submitting data. We are currently developing a deposition tool that will facilitate these submissions and are encouraging others to submit their data.

Our second free resource, the Madison Metabolomics Consortium Database (MMCD, available at www.nmrfam.wisc.edu), acts as a hub for information on biologically relevant small molecules. The MMCD contains the molecular structure, monoisotopic masses, predicted chemical shifts and links for more than 10,000 small molecules. The interface supports single and batch-mode searches by name, molecular structure, NMR chemical shifts, monoisotopic mass, plus various miscellaneous parameters. The MMCD is intended to be a practical tool to aid in identifying metabolites present in complex mixtures.

Another impediment in metabolomics research is the complex logistics associated with coordinating multiple investigators in studies with large numbers of samples. To address this problem, we have created a metabolomics module for our Sesame laboratory information management system (LIMS) [9].

We designed Sesame to capture the complete range of experimental protocols, background information, and experimental data associated with samples. The system allows users to define the actions and protocols to be tracked and supports bar coded samples. Sesame is freely available at http://www.sesame.wisc.edu.

In this paper we discuss the construction and mechanics of these resources as well as the details of our experimental designs and the sources we have drawn upon in developing these tools.

2. Data Model for Metabolomics

The Metabolomics Standards Initiative recently recommended that metabolomics studies should report the details of study design, metadata, experimental, analytical, data processing, and statistical techniques used [10]. Capturing these details is imperative, because they can play a major role in data interpretation [11–13]. As a result, informatics resources need to be built on a data model that can capture all of the relevant information while maintaining sufficient flexibility for future development and integration into other resources [14]. To meet this challenge, the Madison Metabolomics Consortium has adopted a Self-defining Text Archival and Retrieval (STAR) [15–17] for storing and disseminating data.

A STAR file is a flat text file with a simple format and extensible Data Definition Language (DDL). Data are stored as tag-value pairs and loop constructs resemble data tables. The STAR DDL is inherently a database schema that can be mapped one-to-one to a relational database model. Translating between STAR and other exchange file formats, such as XML, is a straightforward process. The STAR DLL used in our metabolomics resources was adapted from the existing data dictionary developed by the BMRB (NMR-STAR) for their work on NMR spectroscopic data of biological macromolecules and ligands.

To describe the data for metabolic standard compounds, we used a subset of the NMR-STAR dictionary suitable for data from small molecules and extended the dictionary to include MS information. The information defined includes a complete compound chemical description (atoms, bonds, charge, etc.), nomenclature (including INChI and SMILES codes and synonyms), mono-isotopic masses, links to databases through accession codes (PubChem, KEGG, CAS, and others), and additional information. Descriptions are provided for the NMR and mass spectrometers and chromatographic systems used in data

collection. Information on the sample contents and sample conditions is captured. Details of the NMR and mass spectrometry experiments can be included. For NMR, pointers to the raw NMR spectral data and the acquisition and processing parameters, experimental spectral peak parameters (peak chemical shifts, coupling constants, line widths, assigned chemical shifts, etc.), chemical shift referencing methods, theoretical chemical shift assignments and details of the calculation methods are described. For MS, the chromatographic retention times for the compound(s) of interest and standards are defined as well as the m/z values and intensities and pointers to the raw data files. The metabolite data dictionary is now being used to construct files containing all of the above information for the growing list of standard metabolic compounds analyzed by our consortium. The populated metabolite STAR files and the raw NMR and MS data files (instrumental binary formats) are being made freely available on the World Wide Web. The BMRB provides tools for converting NMR-STAR files into a relational database and XML files.

3. Metabolite Database at BMRB

3.1. *Approach*

The metabolomics community would clearly benefit from an extensive, freely-accessible spectral library of metabolite standards collected under standardized conditions. Although the METLIN database serves this role for the MS community (http://metlin.scripps.edu/about.php), most current NMR resources have limitations in that they do not provide original spectral data (Sadtler Index [18], NMRShiftDB [19]; NMR metabolomics database of Linkoping (MDL http://www.liu.se/hu/mdl/main/), contain data that were collected under non-standardized conditions ([19], MDL), or do not make their data freely available (AMIX/SBASE http://bruker-biospin.de). To our knowledge, the Human Metabolome Project (http://www.hmdb.ca/) is the only NMR resource, apart from BMRB, without these limitations. The current sparse coverage of NMR metabolomics resources stems in part from the high investment required to compile a comprehensive library of biologically relevant small molecules under standardized conditions. Our solution is to provide at BMRB a well-defined, curated platform that will allow the deposition of data from multiple research groups and free access to all.

3.2. *Rationale for Metabolomics at BMRB*

The BMRB is a logical host for a metabolomics spectral library because of its history as a world wide repository for biological macromolecule NMR data [20–22]. BMRB is a public domain service and is a member of the Worldwide Protein Data Bank. Along with its home office in Madison, Wisconsin, BMRB has mirror sites in Osaka, Japan and Florence, Italy. BMRB is funded by the National Library of Medicine, U.S. National Institutes of Health, and its activities are monitored by an international advisory board. BMRB data are well archived with daily onsite tape backups and offsite third party data backup.

3.3. *Data Collection and Organization*

Currently, the BMRB metabolomics archive contains experimental NMR data for more than 270 compounds collected by the Madison Metabolomics Consortium. Entries contain NMR time-domain data, peak lists, processed spectra, and data acquisition and processing files for one-dimensional (^{1}H, ^{13}C, ^{13}C DEPT 90^{o}, and ^{13}C DEPT 135^{o}) and two-dimensional (^{1}H-^{1}H TOCSY and ^{1}H-^{13}C HSQC) NMR experiments.

A BMRB entry represents a set of either experimental or theoretical data reported for a metabolic compound, mixture of compounds, or experimental sample by a depositor. Entries are further distinguished by the experimental method used (NMR or MS). Separate prefixes on entries serve to discriminate between experimental data (bmse–) and theoretical calculations (bmst–). As described above, the metadata describing the chemical compounds and experimental details and quantitative data extracted from experiments or theoretical calculations for a unique entry are archived in NMR-STAR formatted text files.

On the BMRB ftp site (ftp://ftp.bmrb.wisc.edu/pub/metabolomics), directories are defined for each compound or non-interconverting form of a compound (i.e., L-amino acids). Subdirectories for NMR, MS, and literature data are listed under each compound directory. All data associated with a BMRB experimental or theoretical entry are grouped together in a subdirectory, with the BMRB identifier located under the directory named for the compound studied and the appropriate subdirectory (NMR or MS). Data for compounds that form racemic mixtures in solution (e.g., many sugars) are grouped under a generic compound name.

BMRB has developed internal tools to coordinately view spectra, peak lists, and the molecular structure; these tools are used to review deposited data for

quality assurance purposes. However, the depositor is ultimately responsible for the data submitted, and user feedback is the best defense against erroneous data on a public database. Users who encounter questionable data are encouraged to contact webmaster@bmrb.wisc.edu. Questionable data will be reviewed and corrected if possible; otherwise they may be removed from the site.

3.4. *Presentation and Website Design*

The BMRB metabolomics website has been developed to meet needs expressed by many of its users. The layout and usage of the metabolomics web pages have had several public incarnations and will probably undergo more as the site matures and grows. The first page a visitor sees contains a two-paragraph introduction to the field and a collection of Internet links to a few important small molecule sites; a more complete listing of metabolomics websites is accessed from a link in the sidebar. The information contained in these websites and databases is complementary to that collected by BMRB. The Standard Compounds page (Figure 1) provides the means for searching for metabolites of interest.

For each compound archived, an individual summary page (Figure 2) is created dynamically from the collection of files located in the standard substance sub-directory associated with that compound. A basic chemical description is provided from information BMRB collects from PubChem at the National Institutes of Health, National Center for Biotechnology Information, National Library of Medicine, (http://www.ncbi.nlm.nih.gov/) A two-dimensional stick drawing is created. Three-dimensional '.mol' files are generated from the two dimensional '.sdf' files obtained from PubChem, and these are displayed using Jmol. Links are created to one or more PubChem entries and to the KEGG entry if available. Synonym information and various nomenclature descriptions such as INChI codes, IUPAC names, and SMILES strings are given.

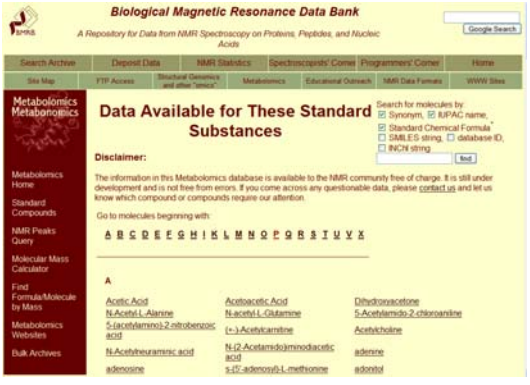

Figure 1. Metabolomics standard substances page in the BMRB website.

The use of dynamic information presentation techniques allows BMRB to create tools that search through the data or calculate answers according to specific user input. NMR data can be displayed in a variety of ways: as a collection of spectra, as a spectrum along with its peak list, or simply a single spectrum of interest. Links allow the user to access the time-domain or processed data by FTP. The Peak Query tool allows the user to enter a list of peaks in one- or two-dimensional formats with tolerances and retrieve a list of compounds with matching signals.

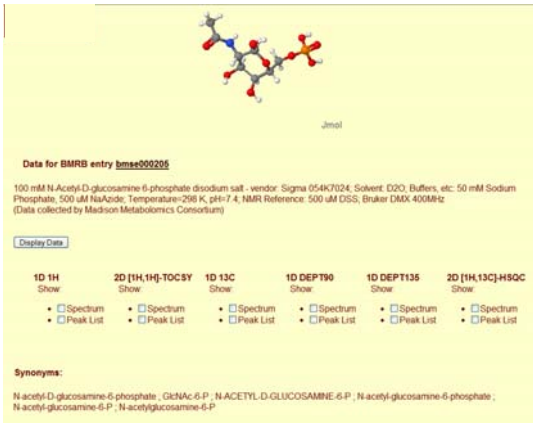

Figure 2. Portion of the substance summary page in the BMRB metabolomics website for N-acetyl-D-glucosamine-6-phosphate.

Historically, multiple approaches have been taken to name or identify small molecules in a standard, unique manner. For this reason, BMRB provides the ability to search on common names, INChI codes, IUPAC names, SMILES strings, and various database identifiers. BMRB allows users to select categories for searching. For example, a user wishing to see all entries for molecules containing nitrogen would search for 'N' with only chemical formula selected. But to find molecules with similar substructures, the INChI or SMILES searches would be the approach to use.

A number of users have requested that the data in the BMRB archive be available through bulk transactions. To accommodate this, we have taken the data from the ftp repository and collated them into a collection of tar (tape archive) files that can be easily downloaded from the ftp site.

3.5. *Prospects*

Over the past year, BMRB with assistance from a grant from the NIH Roadmap Initiative has developed a usable and maintainable metabolomics resource for the research community. A core set of metabolite data from pure compounds has been deposited, and additional data sets are being solicited from the community.

From this beginning, it is clear that enhancements and implementations are needed. A standardized protocol for archiving and presenting mass spectrometry data must be developed. A standard deposition system needs to be completed and brought online, and more efficient methods for validating deposited data need to be put in place. The STAR file DDL needs to be finalized. In addition, it will be useful to develop models and tools that enable users to explore dynamic and parametric relationships among sets of deposited or experimental data.

4. Madison Metabolomics Consortium Database (MMCD)

4.1. *Overview*

The purpose of the MMCD is to be a resource for MS and NMR based metabolomics by providing tools for metabolite identification and characterization. The MMCD was created initially for in-house use, but later was released to the public as hosted on the NMRFAM website (http://www.nmrfam.wisc.edu). The core design features of the MMCD have been practicality and efficiency.

MMCD collects, organizes, and edits information about metabolites from a number of sites, including BMRB. As a starting set, information for 10,912 biologically relevant metabolites was compiled from the KEGG and AraCyc pathway databases, and this list of compounds was supplemented by chemical information obtained from ChemIDplus at the National Library of Medicine Specialized Information Services (http://chem.sis.nlm.nih.gov/chemidplus/) and PubChem. Empirical chemical shifts were predicted for each of the compounds with ChemDraw software. The database also contains theoretical and experimental chemical shifts from BMRB and other sources.

4.2. *Data Content*

Table 1. Three categories of data in the MMCD.

1. Data related to NMR spectroscopy.
Experimental data collected under standard conditions.
Literature data (from NMRShiftDB, etc.).
Chemical shifts from theoretical calculations (by Gaussian 03).
Empirically predicted chemical shifts.
2. Data related to mass spectrometery.
Isotopomer masses for: $^{12}C^{14}N$ / $^{13}C^{14}N$ / $^{12}C^{15}N$ / $^{13}C^{15}N$.
LC-MS data collected under defined conditions.
3. Links to chemical and biological informatics databases: e.g., PubChem; ChemIDplus; KEGG; CHEBI; HMDB and NMRShiftDB databases/websites.

Metabolomics studies require substantial informatics support to identify compounds in complex mixtures. The MMCD was designed to link three categories of data (Table 1). The flexible design of MMCD allows the database to modify both its content and informatics tools depending on demands of the metabolomics community.

Each compound in the database can be characterized by more than 50 data elements. The numbers of compounds with various types of associated data are given in Table 2.

Table 2. Current contents of the MMCD.

Total compounds	10,912
Compounds with experimental NMR data	324
Compounds with theoretical NMR data	150
Compounds with NMR data from the literature	1,000
Compounds with empirical NMR data	10,912

4.3. *Query Engine*

The MMCD has a flexible and efficient query system that allows searches by text, molecular structure, NMR parameters, mass spectrometry parameters (mass, retention time), and miscellaneous other criteria. With its WYSIWYG (what you see is what you get) interface, the user can combine up to five different types of search criteria. The query engine can also be used in batch mode for high-throughput searching.

Clicking on a bar (Figure 3) activates the corresponding search section. Multiple search sections can be activated and queried together as a logical 'AND' relationship. Pushing the "reset" button clears all previous input and restores all sections to their original status.

The "Text-based Search" section (Figure 4) is equipped with a flexible, ambiguous search engine for names or synonyms. Names recognized include those in: CAS, Kegg, CQ_ID, Exp_NMR, Pubchem_SID, ChemIDplus_ID, CHEBI_ID. Synonyms include: Common Name; IUPAC name, Beilstein Handbook Reference ID, EINECS ID, NSC ID, and CCRIS ID.

When "Ambiguous search" is checked (Figure 4), the search will consider as a hit any synonym included as part of the very flexible input format.

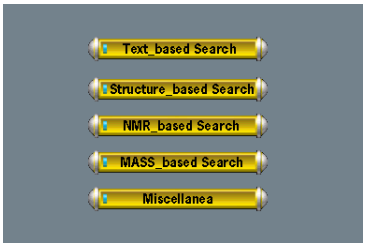

Figure 3. Search criteria in the MMCD.

Wildcards '*' can be used. Two or more names can be submitted separated by ';': e.g. glucose; D-glucose. Such entries are processed according to a logical 'OR' relationship. Salt information is discarded: e.g., gluconic acid sodium salt is considered as gluconic acid. Both *ic acid and *ate yield the same result. e.g., 'D-gluconic acid' is equal to 'D-gluconate'.

Figure 4. Text-based searching of the MMCD.

Clicking on the "Batch mode" bar (Figure 4) enables input from a file. The user needs to identify the file, set the index of the query item (usually begins from 1), and the separator type. Clicking again on the "batch mode" title switches the interface back to the normal search mode.

The other search modules ("Search by Structure", "NMR-Based Search", "Mass-Based Search", and "Miscellanea") have similar functionalities. As noted above, it is possible to combine searches over different criteria. For example one can search on a particular molecular formula ("Structure-based Search") with specified chemical shift values ("NMR-based Search") and tolerances and limit the search to metabolites believed to be associated with *Arabidopsis* ("Miscellanea").

5. Metabolomics Module for the Sesame Laboratory Information Management System

A metabolomics "Module" for the Sesame laboratory information management system (LIMS) has been developed and is being used to organize the activities of the MMC. Sesame is a platform-independent, specialized LIMS written in Java with CORBA used as the middleware. The RDBMS used is Oracle or PostgreSQL, available on multiple platforms. All Sesame Modules contain tools and techniques for collaborative analysis, access, and visualization of data. All Sesame modules are available to the public and are "open source". Sesame is a web-based system that is accessible to team members over the Web. Access is customizable and is password protected. All data are secure and backed up.

The Sesame module for metabolomics, called "Lamp", consists of Small Molecule, Detailed Small Molecule, Sample, Mass Sample, NMR Experiment, Software, Hardware, Vendor and Citation Views and different Lab and System Resources (Lab Protocol, Type, Status, Lock Solvent, Internal Reference Compound, Mass Spectrum Type, Mass Spectrometer, etc.). Views operate on various kinds of data, and facilitate data capture, editing, processing, analysis,

retrieval, or report generation. The data records from different Views can be linked to each other: for example, a Small Molecule entry is linked to a Sample, which is linked to a Mass Sample, NMR Experiment, and a Vendor. Correspondingly, an NMR Experiment is linked to Hardware and Software, etc. Every View contains general and view-specific fields. General fields include the lab name, the date the record was created, the date it was last modified, the lab protocol, user label, information, actions, linked items, attached files, and attached images. View-specific fields include the sample location (room, freezer, tower, box, position in box), description of the observed sample (constituent, name, concentration, concentration unit, isotopic labeling), lock solvent, ionic strength, pH, molecular weight, mass spectrum type, buffer and salt concentration, measured mass, clean-up steps, NMR spectrometer and probe, NMR experiment type, vendor name, address and contact info, etc. The Detailed Small Molecule View contains all the information loaded from different data sources and public databases: e.g., calculated chemical shifts, a two-dimensional image of the molecule drawn from a mol file (if it exists), linked items (samples, mass samples, NMR experiments, vendors, etc.). The Small Molecule View is designed to display results of different queries in a compound fashion. The columns are subsets of the Detailed Small Molecule View: Sesame id, PubChem id, name, formula, weight, 2D image, etc. The data in different views can be queried based on different ids, content, location, status, type, lab protocol used, actions performed, etc. The Lamp module supports queries and reports and export functions.

All standard compounds and experimental samples can be bar-coded and logged into the system. This makes it possible to track the origin, location, amount, and history of each. Experimental data in Sesame can be associated with protocols entered in the system. This allows results on a given tissue to be associated with defined protocols (e.g., for extraction or fractionation).

Acknowledgments

Supported by NIH grant R21 DK070297; I.A.L. is the recipient of a fellowship from the NHGRI 1T32HG002760; NMR data were collected at the National Magnetic Resonance Facility at Madison (NMRFAM) funded by NIH grants (P41 RR02301 and P41 GM GM66326); metabolite standards data are archived at the Biological Magnetic Resonance Data Bank (BMRB), which is supported by a grant from the National Library of Medicine (P41 LM05799).

References

1. I. Nobeli and J. M. Thornton, *Bioessays* **28**, 534 (2006).
2. S. Rochfort, *J. Nat. Prod.* **68**, 1813 (2005).
3. Z. N.Oltvai and A. L. Barabasi, *Science.* **298**, 763 (2002).
4. D. B. Kell, *Curr. Opin. Microbiol.* **7**, 296 (2004).
5. O. Fiehn, *Plant Molec. Biol.* **48**, 155 (2002).
6. J. K. Nicholson, J. Connelly, J. C. Lindon and E. Holmes, E. *Nat. Rev. Drug Discov.* **1**, 153 (2002).
7. W. Weckwerth, *Annu. Rev. Plant. Biol.* **54**, 669 (2003).
8. J. C. Lindon, E. Holmes and J. K. Nicholson, *Prog. NMR Spectr.* **39**, 1 (2001).
9. Z. Zolnai, P. T. Lee, J. Li, M. R. Chapman, C. S. Newman, G. N. Phillips, Jr., I. Rayment, E. L. Ulrich, B. F. Volkman and J. L. Markley, *J. Struct. Funct. Genom.* **4**, 11 (2003).
10. J. C. Lindon et al., *Nature Biotechnol.* **23**, 833 (2005).
11. A. L. Castle, O. Fiehn, R. Kaddurah-Daouk, and J. C. Lindon, *Brief Bioinform.* **7**, 159 (2006).
12. H. Jenkins, N. Hardy, M. Beckmann, J. Draper, A. R. Smith, J. Taylor, O. Fiehn, R. Goodacre, R. J. Bino, R. Hall, J. Kopka, G. A. Lane, B. M. Lange, J. R. Liu, P. Mendes, B. J. Nikolau, S. G. Oliver, N. W. Paton, S. Rhee, U. Roessner-Tunali, K. Saito, J. Smedsgaard, L. W. Sumner, T. Wang, S. Walsh, E. S. Wurtele and D. B. Kell, *Nat. Biotechnol.* **22**, 1601 (2004).
13. E. J. Want, G. O'Maille, C. A. Smith, T. R. Brandon, W. Uritboonthai, C. Qin, S. A. Trauger and G. Siuzdak., *Anal. Chem.* **78**, 743 (2006).
14. D. V. Rubtsov, H. Jenkins, C. Ludwig, J. Easton, M. Viant, U. Guenther, N. Hardy and J. L. Griffin, Proceedings of the 2nd Scientific Meeting of the Metabolomics Society, Boston, Abs. 10 (2006).
15. S. R. Hall, *J. Chem. Inf. Comput. Sci.* **31**, 326 (1991).
16. S. R. Hall and A. P. F. Cook, *J. Chem. Inf. Comput. Sci.* **35**, 819 (1995).
17. S. R. Hall and N. Spadaccini, *J. Chem. Inf. Comput. Sci.* **34**, 505 (1994).
18. *The Sadtler Standard Spectra N.M.R. Chemical Shift Index*, Sadtler Research Laboratories, Philadelphia, (1967).
19. C. Steinbeck, S. Krause and S. Kuhn, *J. Chem. Inf. Comput. Sci.* **43**, 1733 (2003).
20. E. L. Ulrich, J. L. Markley and Y. Kyogoku, *Protein Seq. Data Anal.* **2**, 23 (1989).
21. B. R. Seavey, E. A. Farr, W. M. Westler and J. L. Markley, *J. Biomol. NMR.* **1**, 217 (1991).
22. J. F. Doreleijers, S. Mading, D. Maziuk, K. Sojourner, L. Yin, J. Zhu, J. L. Markley and E. L. Ulrich, *J. Biomol. NMR.* 26, 139 (2003).