Group, during its meeting held on October 25, 2007, to take on board the issues of environmental safety assessment. Rapid and decisive action and definitive proposals are now urgently needed. The alternative—placing international trade in jeopardy—is simply not an option.

Roger Krueger & Bernard Le Buanec

International Seed Federation, Chemin du Reposoir 7, 1260 Nyon, Switzerland.
e-mail: isf@worldseed.org/

1. Organisation for Economic Cooperation and Development. *Recombinant DNA Safety Considerations.* (OECD Publications Service, Paris, 1986).
2. Codex Alimentarius. Guideline for the Conduct of Food Safety Assessment of Foods Derived from Recombinant DNA Plants. CAC/GL 45-2003 (Codex, Rome, 2003)
3. http://www.un.org/millennium/law/cartagena.htm

Metabolite identification via the Madison Metabolomics Consortium Database

To the editor:
High-throughput metabolic profiling, known as metabolomics[1] or metabonomics[2], has been an active area of research for over 35 years[3]. The most commonly employed analytical tools, nuclear magnetic resonance (NMR)[4] and mass spectrometry (MS)[5], have been used extensively to study metabolites in a wide range of biological systems. Despite this long history, bioinformatics resources for identifying common metabolites from experimental NMR and MS data are limited[6]. To remedy this problem, we have developed the Madison Metabolomics Consortium (MMC) Database (MMCD; http://mmcd.nmrfam.wisc.edu/), a web-based tool that contains data pertaining to biologically relevant small molecules from a variety of species.

Identifying metabolites present in unfractionated biological samples is a fundamentally different task from the process of identifying novel natural products. The key distinctions are that the molecular structures of common metabolites are already known, and pure small-molecule standards are often commercially available. This means that many of the time-consuming steps required for natural product compound identification can be replaced by bioinformatics. Efficient bioinformatics-based identifications are critical because of the large number of metabolites that must be evaluated in any given metabolomics study.

Bioinformatics methods for identifying metabolites require an extensive library of experimental data. Most existing libraries (**Supplementary Table 1** online) are either proprietary, insufficiently comprehensive, collected under nonstandardized conditions or unsearchable by computers. Notable exceptions include data in the Human Metabolome Database (HMDB; http://www.hmdb.ca/)[7] and the data collected by the MMC available from the Biological Magnetic Resonance Data Bank (BMRB; http://www.bmrb.wisc.edu/) and the MMCD[6]. These data sets have been collected using standardized protocols relevant to metabolomics.

The MMCD[6] contains information on >20,000 metabolites and other small molecules of biological interest (**Supplementary Table 2** online). These molecules, which were chosen from entries in such databases as KEGG, BioCyc, CHEBI, HMDB, UM-BBD and PDB, represent a collection of primary and secondary metabolites, xenobiotics and common small-molecule contaminants. A total of 477 small-molecule entries in the MMCD contain experimental NMR data collected by the MMC and an additional 525 compounds contain links to NMR data collected by the HMDB. The MMCD and HMDB NMR data have 239 compounds in common. Although the HMDB and MMC collect data under different conditions (HMDB, H_2O, 50 mM phosphate buffer, pH 7.0; MMC, 99.9% D_2O, containing 50 mM phosphate buffer, pH 7.4), the chemical shifts for compounds common to the two are in good general agreement, with an average variation of 0.05 p.p.m. for ^1H chemical shifts and 0.15 p.p.m. for ^{13}C chemical shifts.

The MMCD is more than a data repository. It is equipped with a flexible and efficient query system (**Supplementary Fig. 1** online) that supports complex queries from any combination of its five basic search engines: text, structure, NMR, mass and miscellanea (**Supplementary Methods** and **Supplementary Tutorial** online). Results returned give users access to all of the MMCD information about a molecule and offer direct

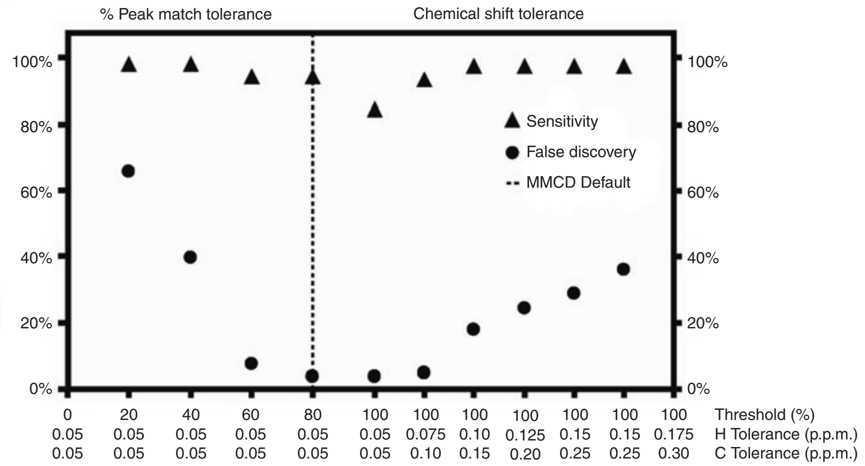

Figure 1 Sensitivity and false discovery rates of the MMCD for batch analysis of mixtures using ^1H-^{13}C HSQC NMR data as a function of the user-controllable peak-matching thresholds. The metabolite mixtures used in this analysis were prepared from pure compounds under the standard MMCD conditions (50 mM phosphate buffer; 99.9% D_2O; pH 7.40, glass electrode reading). Sensitivity (number of correct ID/actual composition) and false discovery (number of incorrect IDs/total metabolites returned) rates were determined as a function of the user-definable ^1H and ^{13}C chemical shift tolerances and peak-match thresholds (the peak-matching threshold defines the number of peaks as a percentage that must be observed in the experimental data for a metabolite to be counted as a positive identification). As expected, tightening the chemical shift and peak-matching tolerances decreased false discovery and reduced sensitivity; conversely, loosening the default thresholds increased both sensitivity and false discovery.

links to records in other public databases, such as the HMDB.

Text-based searches locate metabolites by name or by ID numbers used by other public databases (e.g., KEGG and CAS). Ambiguous and wildcard matches return possible alternatives in cases of misspelled names. Extensive MMCD synonyms lists support multiple molecular naming conventions. Queries can be typed into the graphical user interface or files can be uploaded for batch searching.

The structure-based search engine locates metabolites on the basis of molecular formula, average mass, SMILES (Simplified Molecular Input Line Entry System)[8] string, INCHI (International Chemical Identifiers; http://www.iupac.org/inchi/) string or common structure files (e.g., .mol and .pdb). Alternatively, the structure can be drawn directly into a molecular graphics window. Users can combine as many as six structural criteria in logical/nonlogical fashion to further refine the search and can use controllable similarity thresholds to search for substructures, stereoisomers or related covalent structures.

NMR-based searches give users considerable flexibility with regard to the type and quality of data entered. Chemical shifts can be combined with filters that search for complex multinuclear spin topologies. For example, users can specify chemical shift and atom connectivities (e.g., number of hydrogens attached to a carbon atom). Batch-mode searches return probabilistic identifications of metabolites in mixtures on the basis of various types of one-dimensional (1D) and two-dimensional (2D) NMR data including: 1D ^{1}H, 1D ^{13}C, 2D ^{1}H-^{1}H-TOCSY (total correlation spectroscopy), 2D ^{1}H-^{13}C HSQC (heteronuclear single-quantum coherence), 2D ^{1}H-^{13}C HMBC (heteronuclear multiple-bond correlation) and 2D HSQC-TOCSY. Peak lists can be typed in manually or files can be uploaded in a variety of the common formats used by NMR spectroscopists. NMR searches can use any one of MMCD's three chemical shift databases: experimental (preferred default), empirically predicted from structure (most extensive) or quantum chemical calculated (of interest to theoreticians and useful for quality control and assignment of experimental data). Search results can be downloaded as a tab-delimited file (Excel spreadsheet type) or viewed directly in the MMCD interface. NMR data and structural information have been seamlessly integrated into the search engine. The search engine makes full use of chemical shift, J-coupling and structure-related information, such as connectivity (atom neighbors). The search engine automatically handles issues related to differences in NMR field strength by storing chemical shifts, J-couplings and peak intensities in a field-independent manner. The NMR search engine reconstructs these parameters at the field strength of the data submitted by the user. Thus, qualitative analyses can be carried out using data from any NMR field strength.

The MMCD's NMR-based compound identification tool matches raw peak lists submitted by the users to experimental data collected on pure compounds. The MMCD makes use of the entire pattern of the peaks and has two adjustable tolerances that determine whether a compound is considered to be 'identified'. One threshold controls the permissible chemical shift differences between experimental peaks in the metabolite mixture and those being matched in the database; the other specifies the percentage of cross-peaks from each molecule that must be matched. Loosening the tolerances increases sensitivity and false discovery, whereas tightening the tolerances decreases false discovery at the expense of sensitivity (**Fig. 1**). On the basis of these results, we chose default tolerances of ± 0.05 p.p.m. chemical shift variation for both ^{1}H and ^{13}C and an 80% peak-matching threshold.

We have tested the MMCD's metabolite analysis tool both on mixtures of pure compounds and on metabolite extracts assigned by hand. In the case of the standard mixtures, four solutions were prepared containing a total of 54 metabolites. 2D ^{1}H-^{13}C HSQC NMR spectra were collected and the observed peaks were submitted to the MMCD for analysis. Under the default tolerances, the MMCD averaged 95% sensitivity (identifications corresponding to the compounds known to be present) and 4% false discovery (incorrect identifications). All of the compounds present in the mixtures corresponded to MMCD entries containing experimental data. Thus, these statistics reflect the maximum achievable performance of the MMCD.

Identifying compounds in real biological extracts is complicated by signals from compounds not present in the database, overlap between adjacent peaks and changes in NMR peak position resulting from variations in salt concentration and pH. To evaluate the performance of the MMCD tool on biological samples, we analyzed 2D ^{1}H-^{13}C HSQC spectra of *Arabidopsis thaliana, Medicago truncatula* and *Saccharomyces cerevisiae* extracts that had been assigned by hand[9]. A comparison of metabolite assignments generated by hand and those generated by the MMCD showed that the MMCD produces 45–65% sensitivity (percentage of compounds in agreement with hand assignments) and 0–18% false discovery (assignments that disagreed with those determined by hand).

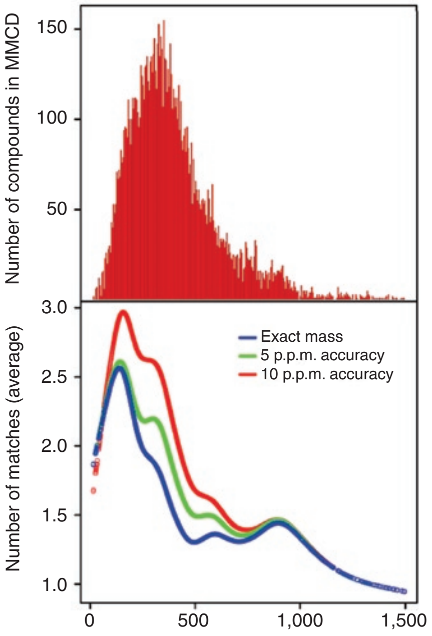

Figure 2 The MMCD mass search engine. (**a**) Distribution of monoisotopic masses (*N* = 19,711) in the MMCD. (**b**) The average number of metabolites returned from a monoisotopic mass query as a function of mass and mass accuracy at 0 p.p.m., 5 p.p.m. and 10 p.p.m. mass accuracy. Results were smoothed (R command smooth.spline; *df* = 18) to determine the average number of matches across a mass range of ~25 p.p.m.

Although the automated approach saves time, the results need to be hand verified. Error checking can be done efficiently by overlaying spectra of the pure standards identified over the experimental spectrum[9]. Once matches have been confirmed, the remaining cross peaks can be submitted to the MMCD for matching, first with the experimental ^{1}H-^{13}C data to find additional matches (at the same or lower tolerances) and then with the database of empirical chemical shifts.

The MMCD mass search engine is primarily designed for identifying metabolites by exact mass (**Fig. 2**), although it can also handle direct-injection MS (DI-MS), liquid chromatography (LC)-MS and MS/MS data. Users can specify the ionization mode, mass accuracy, carbon and nitrogen isotopic composition, and allow for common adducts. Experimental LC-MS and MS/MS peak lists can be uploaded directly either as flat text files or in JCAMP-DX format for batch queries. MS search results can be viewed directly in the MMCD interface or downloaded as a tab-delimited file for viewing with spreadsheet software.

We have also built a 'Miscellanea' search engine that allows users to filter results by the biological species, the type of database to be searched or other criteria. These options allow users to rapidly locate particular sets of records or limit their queries to a preferred subset of records.

In summary, MMCD is a practical tool for expediting the time-consuming steps of identifying and researching small molecules. This freely available resource is compatible with both NMR and MS data (singly or in combination) and facilitates high-throughput metabolomics investigations. Ongoing MMCD support is provided by the National Magnetic Resonance Facility at Madison. Users are encouraged to submit data to the BMRB (supported by the National Library of Medicine, Bethesda, MD, USA), which maintains one of the growing data archives that the MMCD relies upon[10].

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grants R21 DK070297 (MRS, PI) and P41 RR02301 (J.L.M.). We are indebted to W.W. Cleland, H. Lardy and L. Anderson for access to their chemical compound collections. We thank Joe Porwall from Sigma-Aldrich for assistance in locating compounds, Zsolt Zolnai for development of the Sesame LIMS and Eldon L. Ulrich for help in coordinating MMC activities with the BMRB.

Qiu Cui, Ian A Lewis, Adrian D Hegeman, Mark E Anderson, Jing Li, Christopher F Schulte, William M Westler, Hamid R Eghbalnia, Michael R Sussman, & John L Markley

Department of Biochemistry, University of Wisconsin-Madison, 433 Babcock Drive, Madison, Wisconsin 53706, USA.
e-mail: markley@nmrfam.wisc.edu

1. Mendes, P. *Brief Bioinform.* **3**, 134–145 (2002).
2. Nicholson, J.K., Lindon, J.C. & Holmes, E. *Xenobiotica* **29**, 1181–1189 (1999).
3. Pauling, L., Robinson, A.B., Teranishi, R. & Cary, P. *Proc. Natl Acad. Sci. USA* **68**, 2374–2376 (1971).
4. Viant, M.R., Rosenblum, E.S. & Tieerdema, R.S. *Environ. Sci Technol.* **37**, 4982–4989 (2003).
5. Dettmer, K., Aronov, P.A. & Hammock, B.D. *Mass Spectrom. Rev.* **26**, 51–78 (2006).
6. Markley, J.L. *et al. Pac. Symp. Biocomput.* **12**, 157–168 (2007).
7. Wishart, D.S. *et al. Nucleic Acids Res.* **35**, D521–D526 (2007).
8. Weininger, D., Weininger, A. & Weininger, J.L. *J. Chem. Inf. Comput. Sci.* **29**, 97–101 (1989).
9. Lewis, I.A. *et al. Anal. Chem.* **79**, 9385–9390 (2007).
10. Ulrich, E.L. *et al. Nucl. Acids Res.* **36**, D402–D408 (2008).

Human Proteinpedia enables sharing of human protein data

To the editor:

Proteomic technologies, such as yeast two-hybrid, mass spectrometry (MS), protein/peptide arrays and fluorescence microscopy, yield multi-dimensional data sets, which are often quite large and either not published or published as supplementary information that is not easily searchable. Without a system in place for standardizing and sharing data, it is not fruitful for the biomedical community to contribute these types of data to centralized repositories. Even more difficult is the annotation and display of pertinent information in the context of the corresponding proteins. Wikipedia, an online encyclopedia that anyone can edit, has already proven quite successful[1] and can be used as a model for sharing biological data. However, the need for experimental evidence, data standardization and ownership of data creates scientific obstacles.

Here, we describe Human Proteinpedia (http://www.humanproteinpedia.org/) as a portal that overcomes many of these obstacles to provide an integrated view of the human proteome. Human Proteinpedia also allows users to contribute and edit proteomic data with two significant differences from Wikipedia: first, the contributor is expected to provide experimental evidence for the data annotated; and second, only the original contributor can edit their data.

Human Proteinpedia's annotation system provides investigators with multiple options for contributing data including web forms and annotation servers (**Supplementary Fig. 1** online). Although registration is required to contribute data, anyone can freely access the data in the repository. The web forms simplify submission through the use of pull-down menus for certain data fields and pop-up menus for standardized vocabulary terms. Distributed annotation servers[2] using modified protein DAS (distributed annotation system) protocols developed by us (DAS protocols were originally developed for sharing mRNA and DNA data) permit contributing laboratories to maintain protein annotations locally. All protein annotations are visualized in the context of corresponding proteins in the Human Protein Reference Database (HPRD)[3]. **Figure 1** shows tissue expression data for alpha-2-HS glycoprotein derived from three different types of experiments.

Our unique effort differs significantly from existing repositories, such as PeptideAtlas[4] and PRIDE[5] in several respects. First, most proteomic repositories are restricted to one or two experimental platforms, whereas Human Proteinpedia can accommodate data from diverse platforms, including yeast two-hybrid screens, MS, peptide/protein arrays, immunohistochemistry, western blots, co-immunoprecipitation and fluorescence microscopy–type experiments.

Second, Human Proteinpedia allows contributing laboratories to annotate data pertaining to six features of proteins (post-translational modifications, tissue expression, cell line expression, subcellular localization, enzyme substrates and protein-protein interactions; **Supplementary Fig. 2** online). No existing repository currently permits annotation of all these features in proteins.

Third, all data submitted to Human Proteinpedia are viewable through HPRD in the context of other features of the corresponding proteins. To aid comparison and interpretation, meta-annotations pertaining to samples, method of isolation and experimental platform-specific information are provided (e.g., labeling method, protease used, ionization method, details of primary antibody used).

And fourth, in spite of accommodating multiple data types, the data submission is simplified. This means that a biologist with no technical expertise can login and contribute data.

Thus far, a considerable body of data has been contributed to Human Proteinpedia by the community (see **Table 1**), with a total of >1.8 million peptides and >4 million MS/MS-spectra deposited. The above-mentioned data were derived from 2,695 individual experiments (single experiments are defined as immunohistochemistry performed with a specific antibody, a single MS run or a yeast two-hybrid screen). We have imported MS data from two Human Proteome Organization initiatives, including the human plasma proteome project (HPPP)[6] and the human liver