BRIEFINGS IN BIOINFORMATICS. VOL 8. NO 5. 279–293
Advance Access publication July 11, 2007
doi:10.1093/bib/bbm030

Current Progress in computational metabolomics

David S. Wishart

Submitted: 2nd April 2007; Received (in revised form): 15th June 2007

Abstract

Being a relatively new addition to the 'omics' field, metabolomics is still evolving its own computational infrastructure and assessing its own computational needs. Due to its strong emphasis on chemical information and because of the importance of linking that chemical data to biological consequences, metabolomics must combine elements of traditional bioinformatics with traditional cheminformatics. This is a significant challenge as these two fields have evolved quite separately and require very different computational tools and skill sets. This review is intended to familiarize readers with the field of metabolomics and to outline the needs, the challenges and the recent progress being made in four areas of computational metabolomics: (i) metabolomics databases; (ii) metabolomics LIMS; (iii) spectral analysis tools for metabolomics and (iv) metabolic modeling.

Keywords: *databases; bioinformatics; metabolomics; metabolism; cheminformatics*

INTRODUCTION

Metabolomics is a newly emerging field of 'omics' research concerned with the high-throughput identification and quantification of the small molecule metabolites in the metabolome [1]. The metabolome can be defined as the complete complement of all small molecule (<1500 Da) metabolites found in a specific cell, organ or organism. It is a close counterpart to the genome, the transcriptome and the proteome. Together these four 'omes' constitute the building blocks of systems biology. Metabolomics not only serves as a cornerstone to systems biology, it is beginning to serve as a cornerstone to other fields as well. In particular, because of its unique focus on small molecules and small molecule interactions, metabolomics is finding widespread applications in drug discovery [2, 3], drug assessment [3–6], clinical toxicology [5–7], clinical chemistry [8–10], functional genomics [11] and nutritional genomics [12, 13].

Unlike its more mature 'omics' partners, metabolomics is still evolving some of its basic computational infrastructure [14]. Whereas most data in the field of proteomics, genomics or transcriptomics is readily available and readily analyzed through electronic databases, most metabolomic data is still resident in books, journals and other paper archives. Metabolomics also differs from other 'omics' fields because of its strong emphasis on chemicals and analytical chemistry techniques [(nuclear magnetic resonance) NMR, mass spectrometry and chromatographic separations]. As a result, the analytical software used in metabolomics is fundamentally different from any of the software used in genomics, proteomics or transcriptomics. Metabolomics is not only concerned with the identification and quantification of metabolites, it is also concerned with relating metabolite data to biology and metabolism. As a result, metabolomics requires that whatever chemical information it generates must be linked to both biochemical causes and physiological consequences. This means that metabolomics must combine two very different fields of informatics: bioinformatics and cheminformatics.

Despite these differences, metabolomics still shares many of the same computational needs with genomics, proteomics and transcriptomics. All four 'omics' techniques require electronically accessible and

Corresponding author. David Wishart, Department of Computing Science, Department of Biological Sciences and National Institute for Nanotechnology (NRC-NINT), University of Alberta, Edmonton Alberta, Canada T6G 2E8. Tel: 780-492-0383; Fax: 780-492-1071; E-mail: david.wishart@ualberta.ca

David Wishart, a professor at the University of Alberta, has been involved in bioinformatics research since 1990 and metabolomics research since 1998. He currently directs the Human Metabolome Project—a 3-year long project aimed at identifying, characterizing and quantifying all the small molecule metabolites in the human body.

Table I: Summary of metabolite or metabolic pathway databases

Database name	URL or web address	Comments
KEGG (Kyoto encyclopedia of genes and genomes)	http://www.genome.jp/kegg/	Best known and most complete metabolic pathway database Covers many organisms Small (<15) number of data fields, no biomedical data
MetaCyc (encyclopedia of metabolic pathways)	http://metacyc.org/	Similar to KEGG in coverage, but different emphasis Well referenced Small (<15) number of data fields, no biomedical data
HumanCyc (encylopedia of human metabolic pathways)	http://humancyc.org/	MetaCyc adopted to human metabolism
Reactome (a curated knowledgebase of pathways)	http://www.reactome.org/	Pathway database with more advanced query features Not as complete as KEGG or MetaCyc
Roche applied sciences biochemical pathways chart	http://www.expasy.org/cgi-bin/search-biochem-index	The old metabolism standard (on line)
PUMA2 (Evolutionary analysis of metabolism)	http://compbio.mcs.anl.gov/puma2/cgi-bin/index.cgi	Used for metabolic pathway comparison and genome annotation Requires registration
BRENDA (BRaunschweig ENzyme database)	http://www.brenda.uni-koeln.de/	Enzyme database containing rate constants and some metabolic pathway data
Lipid maps	http://www.lipidmaps.org/	Limited to lipids only (not species specific) Nomenclature standard
Chemicals entities of biological interest (ChEBI)	http://www.ebi.ac.uk/chebi/	Covers metabolites and drugs Focus on ontology and nomenclature not biol.
Nicholson's metabolic minimaps	http://www.tcd.ie/Biochemistry/IUBMB-Nicholson/	Used for teaching (limited coverage)

searchable databases, all of them require software to handle or process data from their own high-throughput instruments (DNA sequencers for genomics, microarrays for transcriptomics, mass spectra (MS) for proteomics), all of them require laboratory information management systems (LIMS) to manage their data, and all require software tools to predict or model properties, pathways, relationships and processes.

This review is intended to familiarize readers with the field of computational metabolomics and to highlight the similarities, differences and areas of convergence between metabolomics, genomics, proteomics and transcriptomics. It also outlines the needs and recent progress being made in four key areas of computational metabolomics: (i) metabolomics databases; (ii) metabolomics LIMS and data standards; (iii) spectral analysis tools for metabolomics and (iv) metabolic modeling.

METABOLOMICS DATABASES

Most biochemists and bioinformaticians are familiar with such metabolite and metabolic pathway resources such as KEGG [15], MetaCyc [16] and Reactome [17] along with many others listed in Table 1. These databases, which contain hundreds of reactions, metabolites and pathways for dozens of different organisms, are designed to facilitate the exploration of metabolism and metabolites across many different species. This broad, multi-organism perspective has been critical to enhancing our basic understanding of metabolism and our appreciation of biological diversity. Metabolic pathway databases also serve as the backbone to facilitate many practical applications in biology including comparative genomics and targeted genome annotation. However, the information contained in these 'traditional' databases does not meet the unique data requirements for most metabolomics researchers.

This is because metabolomics is concerned with rapidly characterizing dozens of metabolites at a time and then using these metabolites or combinations of metabolites to identify disease biomarkers or model large-scale metabolic processes. As a result, metabolomics researchers need databases that can be searched not just by pathways or compound

Table 2: Summary of metabolomic databases

Database name	URL or web address	Comments
Human metabolome database	http://www.hmdb.ca	Largest and most complete of its kind. Specific to humans only
BioMagResBank (BMRB – metabolimics)	http://www.bmrb.wisc.edu/metabolomics/	Emphasis on NMR data, no biological or biochemical data Specific to plants (Arabadopsis)
BiGG (database of biochemical, genetic and genomic metabolic network reconstructions)	http://bigg.ucsd.edu/home.pl	Database of human, yeast and bacterial metabolites, pathways and reactions as well as SBML reconstructions for metabolic modeling
Fiehn metabolome database	http://fiehnlab.ucdavis.edu/compounds/	Tabular list of ID'd metabolites with images, synonyms and KEGG links
Golm metabolome database	http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html	Emphasis on MS or GC–MS data only No biological data Few data fields Specific to plants
METLIN metabolite database	http://metlin.scripps.edu/	Human specific Mixes drugs, drug metabolites together Name, structure, ID only
NIST spectral database	http://webbook.nist.gov/chemistry/	Spectral database only (NMR, MS, IR) No biological data, little chemical data Not limited to metabolites
Spectral database for organic compounds (SDBS)	http://www.aist.go.jp/RIODB/SDBS/cgi-bin/direct_frame_top.cgi?lang=eng	Spectral database only (NMR, MS, IR) No biological data, little chemical data Not limited to metabolites

names, but also by NMR spectra, MS, gas chromatography—mass spectrometry (GC–MS) retention indices, chemical structures or chemical concentrations. Likewise metabolomics researchers routinely need to search for metabolite properties, tissue/organ locations or metabolite-disease associations. Therefore, metabolomics databases require information not only about compounds and reaction diagrams, but also data about compound concentrations, biofluid or tissue locations, subcellular locations, physical properties, known disease associations, nomenclature, descriptions, enzyme data, mutation data and characteristic MS or NMR spectra. These data need to be readily available, experimentally validated, fully referenced, easily searched, readily interpreted and they need to cover as much of a given organism's metabolome as possible. In other words, metabolomics researchers need a metabolic equivalent to FlyBase [18] or SwissProt [19].

There are now a number of newly emerging metabolomics databases that are starting to address these needs, either in whole or in part. These include the Human Metabolome Database or HMDB [20], the METLIN database [21], the BioMagResBank or BMRB [22], the Golm Metabolome database [23], the BiGG metabolic reconstruction database [24] and the SDBS [25]. Some, like the HMDB, attempt to address all of the earlier-mentioned database needs, while others, such as SBDS or the BMRB tend to focus on the specific need for creating spectral reference spectral libraries. A brief summary of these and other metabolomic databases is provided in Table 2.

The HMDB is particularly notable for its size, breadth and depth of coverage. It contains physicochemcial, spectral, clinical, biochemical and genomic information for essentially all-known human metabolites, including ~2600 endogenous metabolites and ~250 common exogenous metabolites. Its coverage is approximately three times greater than that of KEGG or HumanCyc. Each metabolite entry contains more than 90 different text fields, images or hyperlinks (Table 3) with much of the information gathered manually or through semi-automated text-mining systems like BioSpider [26]. The database, which is 3.5 Gbytes in size, also supports a wide variety of text, chemical formula, mass, chemical structure, MS spectrum, NMR spectrum and sequence searches (Figure 1). While limited only to human metabolites, the content, display and search capabilities of the HMDB would appear to be good models for other species-specific resources.

Table 3: A summary of data fields in each HMDB 'MetaboCard'

Common name	Cellular location
Metabolite description	Biofluid location(s)
Synonyms	Tissue location(s)
IUPAC name	Concentration (normal urine)
Chemical formula	Concentration (normal plasma)
Chemical structure	Concentration (normal CSF)
Molecular weight	Concentration (normal other biofluids)
Smiles string	Associated disorders
KEGG compound ID	Concentration (abnormal urine)
PubChem ID	Concentration (abnormal plasma)
OMIM ID	Concentration (abnormal CSF)
MetaGene ID	Concentration (abnormal other fluids)
ChEBI ID	Pathway names
CAS registry number	Pathway images
InChi identifier	Pathway graphs
Synthesis reference	Pathway SBMLs
Melting point	Metabolic enzyme name
Water solubility (experimental)	Metabolic gene name
Water solubility (theoretical)	Metabolic enzyme synonyms
LogP	Enzyme protein sequence
Compound state (solid, liquid, gas)	Number of residues
MSDS sheet	Molecular weight
MOL image and file	Enzyme theoretical pI
SDF file	Gene ontology classification
PDB file	General function
PDB image	Enzyme pathway
Predicted 1-H NMR spectrum	Enzyme reaction
Predicted 13-C NMR spectrum	Enzyme PFAM domain
Observed 1-H NMR spectrum	Enzyme transmembrane regions
Observed 13-C NMR spectrum	Metabolic importance
EI mass spectrum	Gene sequence
Ion trap mass spectrum	Chromosome location/locus
Related references	SNPs

In addition to the HMDB, several other organism-specific metabolomic databases are also available, including the BMRB (for Arabadopsis), the Golm Database (primarily plants), the BiGG database (for selected bacteria, yeast and humans) and the METLIN database (for humans). The METLIN database contains chemical structure, chemical formula, mass and nomenclature data on more than 15 000 known and hypothesized human metabolites [21]. While this collection covers a significant number of endogenous metabolites, many of the listed compounds are actually drugs or drug metabolites. Additionally, more than 8000 hypothesized di and tripeptides are included in this total. The inclusion of hypothesized metabolites along with drugs or drug metabolites is somewhat questionable, given the usual practices in metabolomics.

A particular challenge for any species-specific database is defining what should be included in the metabolome. Should the metabolome be restricted to only endogenous compounds? Should it include exogenous drugs and drug metabolites? Should it include plant-derived food components or chemical food additives? Is it appropriate to include hypothetical compounds (such as all combinations of di and tripeptides or all-known chemicals that an organism might ever have any contact with)? What is the upper size or upper molecular weight limit for something to be called a metabolite (<1500 daltons)? Should the metabolome be restricted to those compounds that can be practically detected or detectable? Unlike the genome which is a clearly defined entity, the metabolome (like the proteome) has many definitions for many different people. This appears to be a source for both confusion and dispute within the metabolomics community [27, 28]. Hopefully the introduction of data standards [29, 30] and the establishment of dedicated bodies (such as the Metabolomics Society) will go a long way in resolving this issue.

METABOLOMIC LIMS AND DATA STANDARDS

While not as 'sexy' or scientifically challenging as other aspects of computational metabolomics, the management, storage and standardization of metabolomic data is absolutely critical to making metabolomics more fully integrated into the other 'omics' sciences [29]. Similar standardization efforts proved to be critical to the success and growing uniformity of many techniques in genomics, transcriptomics and proteomics [31]. One of the best routes to achieving data standardization is through the development, distribution and widespread use of mark-up languages (XML, CellML, SBML) and bio-ontologies [32, 33]. The use of common languages or common ontologies allows data not only to be more widely exchanged across disciplines but also to be more easily handled by a much wider number of software packages.

Another critical approach to data standardization lies in making instrumental data more uniformly readable and more easily exchanged. Compared to other 'omics' disciplines, metabolomic data can be collected by a much wider variety of instruments

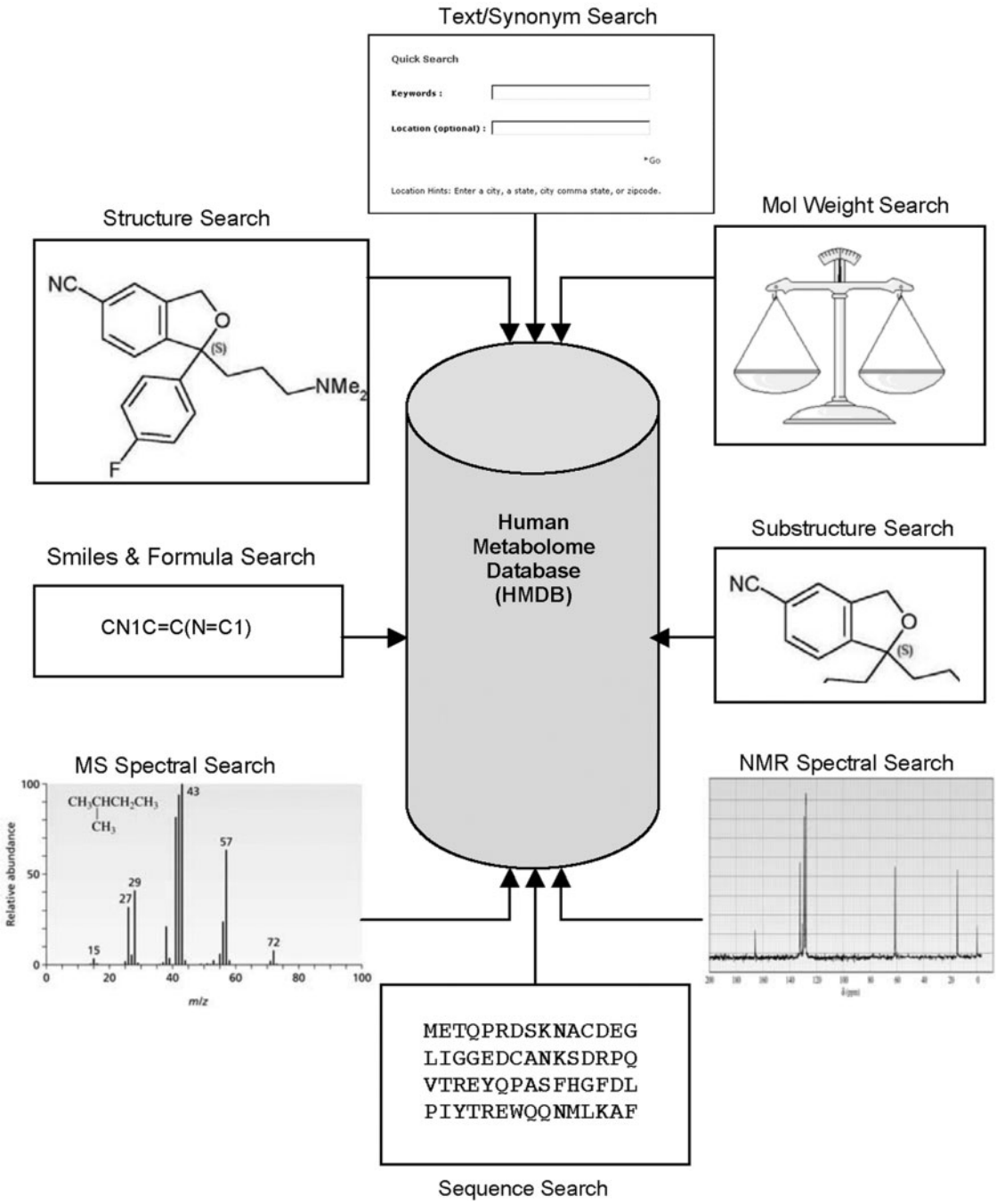

Figure 1: Schematic of the types of search queries supported by the HMDB. Other metabolomics databases are beginning to support similar kinds of search queries.

(HPLC, UPLC, GC–MS, FTIR, LC–MS, NMR) from a much wider range of manufacturers. This can (and has) created a 'Tower of Babel' effect with every metabolomic instrument or laboratory speaking its own unique language. This prevents users from easily converting, normalizing, correcting or aligning their spectral or chromatographic data when working from more than one type of instrument or when working with data measured from other laboratories. Therefore, one of the key challenges in computational metabolomics lies in developing standardized protocols for converting and archiving instrument data to a common format suitable for any kind of mathematical analysis. One possible solution is the adoption of the NetCDF (Network Common Data Form) and ANDI (ANalytical Data

Interchange protocol) file protocols. NetCDF is a general-purpose, machine-independent file protocol for creating, sharing and saving scientific data of almost any kind. It is self-describing, portable, directly accessible, appendable, sharable and archivable. NetCDF was developed by the Unidata Program Center in Boulder, Colorado. On the other hand, ANDI is a more specific file protocol for saving HPLC, UPLC, CE, FTIR and mass spectrometry data. It was originally implemented by the Analytical Instrument Association (AIA) and is supported by a number of instrument manufacturers. More importantly, ANDI is based on the NetCDF file protocol. More details about these file protocols are available at http://www.astm.org/ and http://www.unidata.ucar.edu/

Yet another route to metabolomic data standardization is through the integration of common ontologies, common reporting standards and common data formats into LIMS. A LIMS is computer software system that is used in the laboratory for the management of samples, laboratory users, instruments, standards, workflow automation and other laboratory functions [34]. In other words, LIMS are essentially electronic-record-keeping systems. They are particularly useful for coordinating large-scale, multi-lab or multi-investigator projects and in bringing some semblance of uniformity to input data.

These days, LIMS must adhere to a number of strict data handling requirements. They must allow sample tracking (sample arrival, location, collection data), the storage of methods, protocols and SOPs and the entry of daily lab diaries (a lab notebook). LIMS must also support data time stamps and regular data back up, resource (equipment) and personnel management, data validation, lab audits and the maintenance of lab and data security (an audit trail). Since LIMS are designed to handle large quantities of very heterogeneous data, they have become a mainstay of many 'omics' efforts around the world. Not only are LIMS important for the day-to-day lab management of many of today's large-scale genomic and proteomic experiments, LIMS also play a crucial role in defining what kinds of data will reside in public databases; what kind of data exchange standards will be used for a given field; what kinds of common vocabularies or ontologies should be adopted and what kinds of meta data should be captured during a given experiment. Over the past decade, a number of excellent LIMS have been developed and described for DNA sequencing [35], MS-based proteomics [36], transcriptomics [37] and structural proteomics [38].

While LIMS for genomics, transcriptomics and proteomics have been around for many years, metabolomic LIMS are just beginning to be developed and implemented. Some of the most recent examples include SetupX [39] and Sesame [40]. These metabolomic data management systems build on the experience and conventions established by previous LIMS efforts in genomics, transcriptomics and proteomics.

SetupX, developed by the Fiehn laboratory at UCSD, is an excellent example of a web-based metabolomics LIMS. It is XML compatible and built around a relational database management core. It is particularly oriented towards the capture and display of GC–MS metabolomic data through its metabolic annotation database called BinBase. Under development since 2003, SetupX was originally based on the general 'Architecture for a metabolomics experiment' schema called ArMet [41]. SetupX was designed to be very flexible, being able to handle a wide variety of BioSources (spatial, historical, environmental and genotypic descriptions of biological objects undergoing metabolomic investigations) and Treatments (experimental alterations that influence the metabolic states of BioSources).

A particular strength of SetupX is its use of publicly available taxonomic and ontology repositories to ensure data integrity and logical consistency of its BioSource and Treatment data. For example, selection of BioSource 'rat' and a plant organ 'leaf' is prevented in SetupX. SetupX also uses the NCBI taxonomy tables to enable queries for synonyms or generalized terms such as the genus 'rat' for any of the 23 rat species that are currently defined at the NCBI. Overall, SetupX represents a very flexible, well-designed and well-tested LIM system. It makes use of many leading-edge computational techniques and employs many of the recommendations made by various standing committees on metabolomic data exchange and data entry [29, 30, 41].

Sesame [40] is another example of a web-based, platform-independent metabolomic LIMS. It is written in Java that can use either Oracle or PostgreSQL as its relational database management system (RDBMS). Originally developed to facilitate NMR-based structural genomics studies [42], Sesame is flexible enough to have been recently adapted to handling NMR- (and MS) derived

metabolomic data as well. Like most LIMS, Sesame contains a plethora of tools and techniques to facilitate collaborative analysis, access and visualization of data. Sesame also supports sample tracking and bar coding as well as the entry of standard operating protocols (SOPs) or procedures, such as those that might be used for metabolite extractions or biofluid fractionations.

The Sesame module for metabolomics is called 'Lamp'. The Lamp module was designed primarily for metabolomic studies of Arabadopsis using NMR, although it is flexible enough to be easily adapted to other biological systems and other analytical methods. It consists of a number of different 'Views' each of which provide details about the data, the instruments or system resources used in a given study. These Views or panels cover many of the typical components found in a metabolomics experiment including: Small Molecule, Detailed Small Molecule, Sample, Mass Sample, NMR Experiment, Software, Hardware, Vendor, etc. In Sesame, the Views are designed operate on various kinds of data, and facilitate data capture, editing, processing, analysis, retrieval or report generation.

Overall, the Sesame/Lamp system is a very comprehensive and well-designed LIMS. It has undergone several years of real-world testing and it certainly meets the needs of several different user communities. The Sesame/Lamp system shares a number of features with SetupX (web-enabled, extensive sample tracking, support for metabolite annotation), but differs in its overall design and presentation. This simply underlines the fact that there is no 'right' way to build a LIMS. Likewise, there is no single LIMS that will serve all users. Each LIMS has to be adapted to the needs, preferences and styles of the different labs or different individuals that use them.

SPECTRAL ANALYSIS TOOLS FOR METABOLOMICS

In some respects, metabolomics—or metabolic profiling—is a bit like clinical chemistry. Both are relatively non-invasive diagnostic techniques that look at small molecules from tissues, cells or biofluids. However, what distinguishes metabolomics from clinical chemistry is the fact that in metabolomics one is measuring not just one or two compounds at a time, but literally hundreds at a time. Furthermore, in clinical chemistry, most metabolites are typically identified and quantified using colorimetric chemical assays. In metabolomics, large numbers (tens to hundreds) of metabolites are rapidly (minutes) measured using non-chemical, non-colorimetric methods such as GC–MS, LC–MS (liquid chromatography—mass spectrometry), CE (capillary electrophoresis), FT-MS (Fourier transform mass spectrometry) or NMR spectroscopy [43].

There are two very distinct routes or schools-of-thought for collecting, processing and interpreting metabolomic data (Figure 2). In one version (the chemometric or non-targeted approach), the compounds are not formally identified—only their spectral patterns and intensities are recorded, compared and used to make diagnoses, identify phenotypes or draw conclusions [44, 45]. In the other version (targeted profiling), the compounds are formally identified and quantified. The resulting list of compounds and concentrations (a metabolic profile) is then used to make diagnoses, identify phenotypes or draw conclusions [8, 46].

SPECTRAL ANALYSIS—CHEMOMETRICS AND METABOLOMIC DATA ANALYSIS

Chemometrics can be defined as the application of mathematical, statistical, graphical or symbolic methods to maximize the information which can be extracted from chemical or spectral data. Chemometric approaches for spectral analysis emerged in the 1980s and are primarily used to extract useful information from complex spectra consisting of many hard-to-identify or unknown components [47, 48]. Chemometric approaches can also used to identify statistically significant differences between large groups of spectra collected on different samples or under different conditions.

To facilitate the spectral analysis process, each input spectrum is usually divided up into smaller regions or bins. This spectral partitioning process is called 'binning' and it allows specific features, peaks or peak clusters in a multi-peak spectrum to be isolated or highlighted (Figure 3). Once binned, the peak intensities (or total area under the curve) in each bin are tabulated and analyzed using multivariate statistical analysis. This 'divide-and-conquer' approach allows spectral components to be quantitatively compared within a single spectrum or between multiple spectra. Of course the number

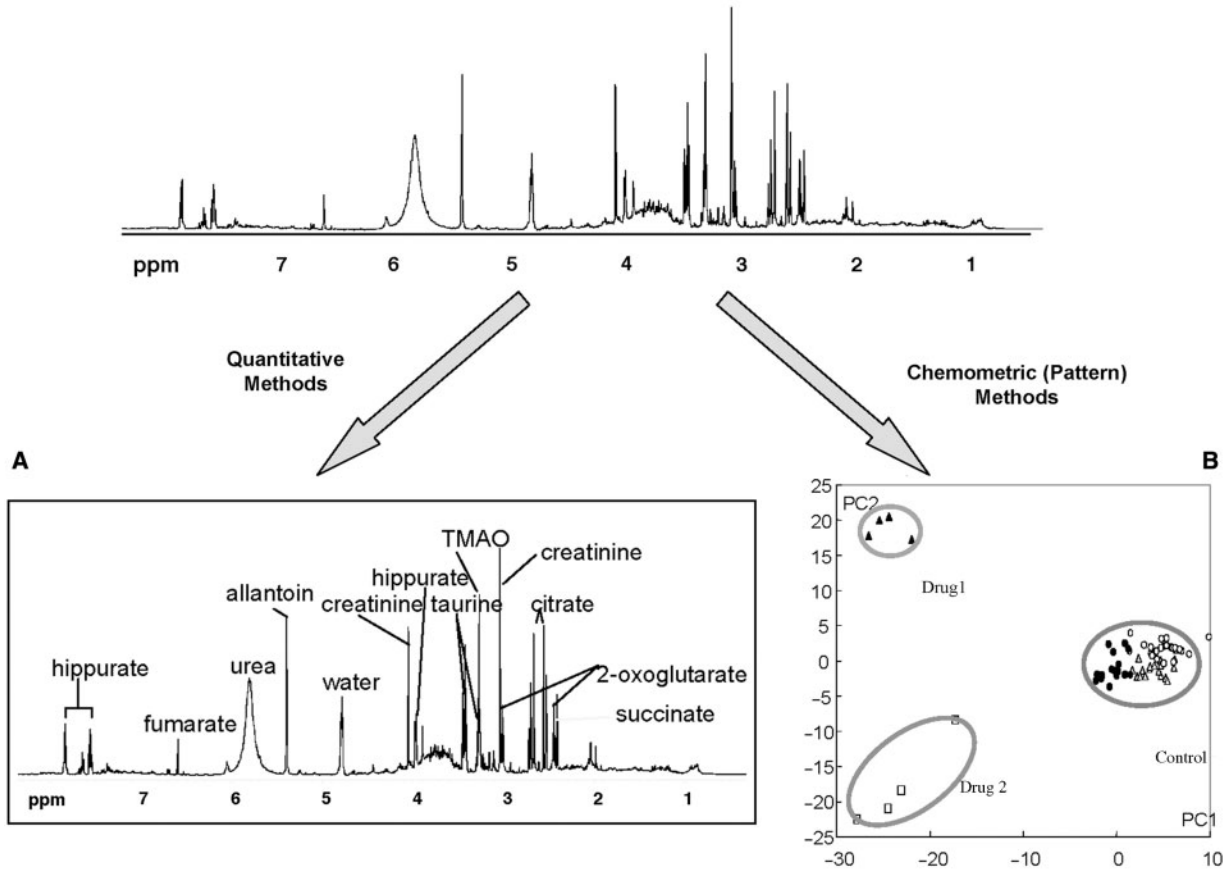

Figure 2: Two approaches to metabolomics: (**A**) Targeted profiling (Bottom-Up) and (**B**) Chemometric (Top-Down). In targeted profiling, the compounds are identified and quantified prior to analysis. In chemometric approaches, compounds are not necessarily identified; only their patterns or peak features are used in data analysis.

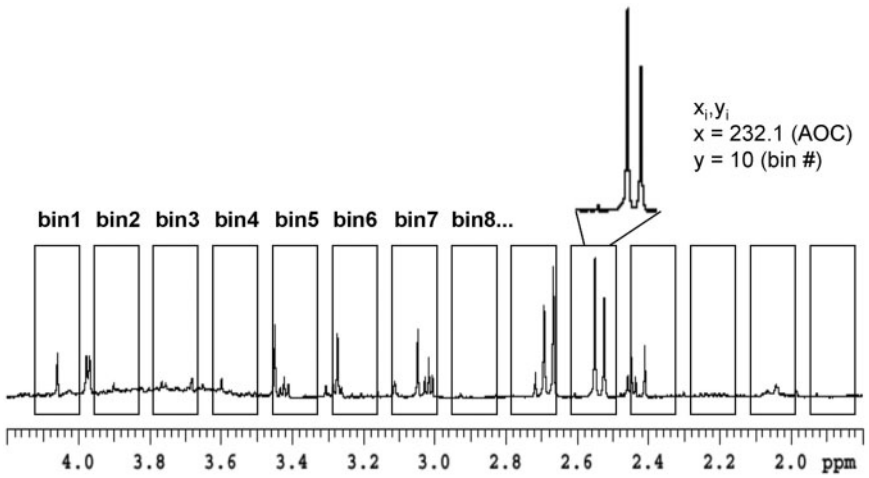

Figure 3: A diagram of spectral binning applied to an NMR spectrum. Spectral binning is one of the prerequisites for PCA.

of components or 'dimensions' that a binned spectrum may represent could number in hundreds or even thousands. To reduce the complexity or the number of parameters, chemometricians use dimensional reduction to identify the key components that seem to contain the maximum amount of information or which yield the greatest differences. The most common form of dimensional

reduction is known as Principal Component Analysis or PCA.

PCA is not a classification technique; rather it is an unsupervised clustering or data reduction technique. Specifically, PCA determines an optimal linear transformation for a collection of data points such that the properties of that sample are most clearly displayed along the coordinate (or principal) axes. PCA is particularly useful to identify how one sample is different from another, which variables contribute most to this difference and whether those variables contribute in the same way (i.e. are correlated) or independently (i.e. uncorrelated) from each other. PCA also quantifies the amount of useful information or signal that is contained in the data. While PCA methods can help quantify information content, they are still quite sensitive to experimental noise. This is because all data dimensions (both metabolite-containing and non-metabolite bins) are typically included in generating the final and reduced models.

As a data reduction technique, PCA is particularly useful as it allows one to easily detect, visually or graphically, sample patterns or groupings. PCA methods can also be extended to higher-order arrays such as three-way data (i.e. data that can be arranged in a cube rather than a table) using a technique called PARAFAC (Parallel Factor Analysis). So while PCA methods work well for analyzing binned NMR, GC–MS or HPLC data (i.e. data with peak height and peak location), PARAFAC methods can be applied to three (and higher) dimensional GC–GC–MS data, 2D–HPLC–MS data or 3D-NMR data.

PCA is not the only chemometric or statistical approach that can be applied to spectral analysis in metabolomics or metabonomics. In fact, there are many other statistical techniques that are available including SIMCA (Soft Independent Modeling of Class Analogy), PLS–DA (Partial Least Squares—Discriminant Analysis) and *k*-means clustering. All of these techniques have been used to interpret NMR, MS/MS and FTIR spectral patterns in a variety of metabolomic or metabonomic applications [49–51].

Like PCA, SIMCA maps its data onto a much lower dimensional subspace for classification. However, unlike PCA, SIMCA uses cross validation or training to take the unlabeled or unidentified PCA clusters and to perform classifications. So in SIMCA, an unknown is only assigned to a class for which it has a high probability. If the residual variance of a sample exceeds the upper limit for every modeled class in the data set, then the sample would not be assigned to any of the existing classes because it is either an outlier or comes from a class that is not represented in the data set. Another advantage to SIMCA is the fact that it is sensitive to the quality of the data used to generate the principal component models. SIMCA techniques in combination with ^{1}H NMR have been used to identify and classify different teas from around the world [52], to classify different types of whiskeys [53] from GC–MS analyses and to perform metabolic phenotyping of nude and normal mice using NMR spectra [54]. Many other examples of SIMCA applications to metabolomics now exist in the literature [44].

PLS–DA is another supervised classification technique, meaning that information about the class identities has to be provided by the user in advance of running the analysis. PLS–DA is used to sharpen the separation between groups of observations, by essentially rotating PCA components such that a maximum separation among classes is obtained. In doing so, it is hoped that one can better understand which variables carry the class-separating information. The principles behind PLS (partial least squares) are similar to that of PCA. However, in PLS, a second piece of information is used, namely, the labeled set of class identities. The PLS algorithm maximizes the covariance between the 'test' or predictor variables and the training variable(s). PLS–DA, which is a particular form of PLS, is a regression or categorical extension of PCA that takes advantage of a priori or user-assigned class information to attempt to maximize the separation between groups of observations. PLS–DA in combination with near infrared spectroscopy has been used to classify the geographic location of wines [55], to look at gender differences in urinary glucuronides via MS-TOF studies [56], and to identify biomarkers in cerebrospinal fluid via SELDI-MS [57].

The intent in using pattern classification for spectral analysis is not to identify any specific compound but, rather, to look at the spectral profiles of biofluids or tissues and to classify them in specific categories, conditions or disease states. This trend to pattern classification represents a significant break from the classical methods of analytical chemistry or traditional clinical chemistry which historically have depended on identifying and quantifying

specific compounds. With chemometric profiling methods one is not so interested in quantifying known metabolites, but rather in trying to look at all the metabolites (known and unknown) at once [44, 45, 48]. The strength of this holistic approach lies in the fact that one is not selectively ignoring or including key metabolic data in making a phenotypic classification or diagnosis. These pattern classification methods can perform quite impressively and a number of groups have reported success in diagnosing certain diseases such as colon cancer [50], in identifying inborn errors of metabolism [8], in monitoring organ rejection [58] and in classifying different strains of mice and rats [51, 59]. A comprehensive discussion of all the chemometric methods being used in metabolomics is beyond the scope of this review. However, for those readers wanting more information about chemometric analyses in metabolomics or more details about the strengths and weaknesses of these approaches, several excellent reviews are now available [44, 47, 60].

SPECTRAL ANALYSIS—TARGETED METABOLIC PROFILING

Targeted metabolic profiling is fundamentally different than most chemometric approaches. In targeted metabolic profiling, the compounds in a given biofluid or tissue extract are actually identified and quantified by comparing the biofluid spectrum of interest to a library of reference spectra of pure compounds [8, 46, 61, 62]. The basic assumption in targeted profiling is that the spectra obtained for the biofluid (which is a mixture of metabolites) is the sum of individual spectra for each of the pure metabolites in the mixture. This approach to compound identification is somewhat similar to the approach historically taken by GC–MS methods and, to a much more limited extent, LC–MS methods [63, 64]. For NMR, this particular approach requires that the sample pH be precisely known or precisely controlled. It also requires the use of sophisticated curve-fitting software and specially prepared databases of NMR spectra of pure metabolites collected at different pH values and different spectrometer frequencies (400, 500, 600, 700 and 800 MHz) [46].

One of the strengths of the NMR-curve fitting approaches is the fact that the NMR spectra for many individual metabolites are often composed of multiple peaks covering a wide range of chemical shifts. This means that most metabolites have unique or characteristic 'chemical shift' fingerprints. This particular characteristic of NMR spectra helps reduce the problem of spectral (or chromatographic) redundancy as it is unlikely that any two compounds will have identical numbers of peaks with identical chemical shifts, identical intensities, identical spin couplings or identical peak shapes. Likewise, with higher magnetic fields (>600 MHz) the chemical shift separation among different peaks and different compounds is often good enough to allow the unambiguous identification of up to 100 compounds at a time—through simple curve fitting [8, 46, 62].

Targeted metabolic profiling is not restricted to NMR or GC–MS. It is also possible to apply the same techniques to LC–MS systems [64]. In the case of MS spectroscopy, the sample MS/MS spectra must be collected at reasonably similar collision energies and on similar kinds of instruments [65]. In other words, an MS/MS fingerprint library determined from a triple-quad instrument will only work with data derived from other triple-quad instruments, while a fingerprint library derived from an ion-trap instrument is specific to the data derived from other ion-trap instruments. Quantification of metabolites by LC–MS is somewhat more difficult than GC–MS or by NMR. Typically quantification requires the addition or spiking of isotopically labeled derivatives of the metabolites of interest to the biofluid or tissue sample. The intensity of the isotopic derivative can then be used to quantify the metabolite of interest.

A key advantage targeted metabolic profiling is that it does not require the collection of identical sets of cells, tissues or lab animals and so it is more amenable to human studies or studies that require less day-to-day monitoring (i.e. no requirement for metabolic chambers). A key disadvantage of this approach is the relatively limited size of most current spectral libraries (~250 compounds). Such a small library of identifiable compounds may bias metabolite identification and interpretation. Both the targeted and chemometric approaches have their advocates. However, it appears that there is a growing trend towards combining the best features of both methods.

Since targeted metabolic profiling yields information about both the identity and concentration

of compounds, it is possible to use a large range of statistical and machine-learning approaches to interpret the data. In fact, the same statistical techniques used in chemometric or non-targeted studies—PCA, SIMCA, PLS–DA, *k*-means clustering—can also be used with targeted profiling. However, instead of using binned spectra or arbitrary peak clusters as input to these algorithms, the actual names of the compounds and their concentrations are used as input. This added specificity seems to significantly improve the discriminatory capabilities of most statistical techniques over what is possible for unlabeled or binned spectral data [46]. Targeted profiling also seems to be particularly amenable to other, more powerful, classification techniques. In particular, the similarity between the information in targeted metabolite profiles (compound names + concentrations) and the information found in microarrays (gene names + transcript abundance) or proteomic profiles (protein names + copy numbers) also means that even more sophisticated machine-learning approaches can be used to analyze this sort of data. These machine-learning approaches include artificial neural networks (ANNs), support vector machines (SVMs) and Decision Trees (DTs).

For instance, ANN analysis of metabolite profiles has been used to identify the mode of action for herbicides on plant biochemical pathways [66]. Using NMR data collected on plant extracts, the authors were able to distinguish or classify the modes of action for 19 different herbicides and identify the metabolic pathways in corn plants that these herbicides affected. The discriminatory power of the ANN method was refined and cross-validated on 400 different plant samples that were treated with many different herbicides.

METABOLIC MODELING AND THE INTERPRETATION OF METABOLOMIC DATA

As we have already seen, the statistical and computational methods described earlier are particularly useful for identifying metabolic differences or finding interesting biomarkers. However these approaches are not designed to provide a great deal of biological insight nor can they provide clear perspective on the underlying biological causes for the metabolic profiles that are seen. To gain this sort of insight, it is often necessary to either mine the literature or to turn to metabolic modeling. Metabolic modeling or metabolic simulation can be done in a variety of ways. Traditionally, it is done by writing down and solving systems of time-dependent ordinary differential equations (ODEs) that describe the chemical reactions and reaction rates of the metabolic system of interest. There are now a host of metabolic simulation programs that allows very complex, multi-component simulations to be performed [67, 68]. These include programs such as GEPASI [69], CellDesigner [70], SCAMP [71] and Cellerator [72]. GEPASI is a good example of a typical metabolic or biochemical pathway simulation package. This program, which has been under development for almost 15 years, uses a simple interface to allow one to build models of metabolic pathways and simulate their dynamics and steady state behavior for given sets of parameters. GEPASI also generates the coefficients of Metabolic Control Analysis for steady states. In addition, it allows one to study the effects of several parameters on the properties of the model pathway. GEPASI can also be used to simulate systems with stable states, limit cycles and chaotic behavior. GEPASI allows users to enter the kinetic equations of interest and their parameters (Km, reaction velocity, starting concentrations), solves the ODEs using an ODE solver and generates plots that can be easily visualized by the user. GEPASI has been used in a wide variety of metabolic studies such as bacterial glucose/galactose metabolism [73] and glutathione/phytochelitin metabolism [74] and continues to be used in many metabolomic or kinetic analyses.

An alternative to solving large systems of time-dependent rate equations is a technique known as constraint-based modeling [75, 76]. Constraint-based modeling uses physicochemical constraints such as mass balance, energy balance and flux limitations to describe the potential behavior of a large metabolic system (a cell, an organ, an organism). In this type of modeling, the time dependence and rate constants can be ignored as one is only interested in finding the steady state conditions that satisfy the physicochemical constraints. Since cells and organs are so inherently complex and because it is almost impossible to know all the rate constants or instantaneous metabolite concentrations at a given time, constraint-based modeling is particularly appealing to those involved in large-scale

metabolomic studies. In particular, through constraint-based modeling, models and experimental data can be more easily reconciled and studied on a whole-cell or genome-scale level [75, 76]. Furthermore, experimental data sets can be examined for their consistency against the underlying biology and chemistry represented in the models.

One of the most popular approaches to constraint-based metabolic modeling is known as flux-balance analysis or FBA [77, 78]. FBA requires knowledge of the stoichiometry of most of reactions and transport processes that are thought to occur in the metabolic system of interest. This collection of reactions defines the metabolic network. FBA assumes that the metabolic network will reach a steady state constrained by stoichiometry of the reactions. Normally the stoichiometric constraints are too few and this leads to more unknowns than equations (i.e. an underdetermined system). However, possible sets of solutions can be found by including information about all feasible metabolite fluxes (metabolites added or excreted) and by specifying maximum and minimum fluxes through any particular reaction. The model can also be refined or further constrained by adding experimental data—either from known physiological or biochemical data obtained from specific metabolomic studies. Once the solution space is defined, the model is refined and its behavior can be studied by optimizing the steady state behavior with respect to some objective function. Typically the objective function optimization involves the maximization of biomass, the maximization of growth rate, the maximization of ATP production, the maximization of the production of a particular product or the maximization of reducing power. Once the model is fully optimized, it is possible to use that FBA model to create predictive models of cellular, organ or organismal metabolism. These predictions can be done by changing the network parameters or flux balance, changing the reactants, adding new components to the model or changing the objective function to be maximized.

Critical to the success of any FBA model is the derivation or compilation of appropriate mass balance [76, 77]. Mass balance is defined in terms of both the flux of metabolites through each reaction and the stoichiometry of that reaction. Mass balance considerations give rise to a set of coupled differential equations. This set of equations is often expressed as a matrix equation which can be solved through simple linear algebra and optimized through linear programming. The goal of FBA is to identify the metabolic fluxes in the steady state (i.e. where the net flux is 0). Since there are always more reactions than metabolites, the steady state solution is always underdetermined. As a result, additional constraints must be added to determine a unique solution. These constraints can be fluxes measured through metabolomics experiments (such as isotope labeling experiments) or through estimated ranges of allowable (feasible) flux values. FBA methods can also incorporate regulatory constraints, explicit incorporation of thermodynamic constraints or different objective functions.

FBA methods have been used in a variety of metabolomic studies. In particular, they have been used in the genome-scale modeling of many bacterial metabolic systems including *Lactococcus lactis, Corynebacterium glutamicum, Streptomyces coliecolor, Helicobacter pylori* and *Escherichia coli* [79–83]. Flux balance analysis has also been used to look at yeast metabolism [84, 85], erythrocyte metabolism [86], myocardial metabolism [87] and most impressively the entire human metabolomic network [24]. Certainly, as more detailed flux data is acquired through isotope tracer analysis and more information is obtained from quantitative, targeted metabolic profiling, it is likely that flux balance analysis and other kinds of constraint-based modeling will play an increasingly important role in the interpretation of metabolomic data—and in computational metabolomics.

CONCLUSIONS

Metabolomics is a very young field and consequently computational metabolomics is even younger. However, by following in the computational footsteps of other 'omics' efforts—and avoiding their pitfalls, it is very likely that the field of computational metabolomics will rapidly catch up. Currently the areas of most active development include the creation of comprehensive metabolomics databases, the establishment of data exchange and data storage standards, the refinement of data analysis tools and the improvements in metabolic modeling. Nevertheless, despite these developments, it is also clear that there are still many opportunities for algorithmic development and bio/chemo- informatics innovation.

Certainly, the continuing development of newer and better technologies for metabolite detection and quantification will continue to drive a significant portion of computational metabolomics, particularly in the areas of data reduction, normalization and alignment. There may also be a gradual consolidation of techniques or technologies, allowing a greater degree of data standardization and an increased level of data sharing among different instruments or different labs. The appearance of completely novel, revolutionary or 'disruptive' technologies such as metabolite microarrays or hand-held metabolite 'tricorders' is not beyond reason and certainly these revolutionary technological changes would have similarly revolutionary consequences for computational metabolomics.

Regardless of the coming technological changes, it is almost certain that computational metabolomics will become increasingly integrated or aligned with systems biology [88]. This integration will require that metabolomics methods and data reduction techniques will have to become much more quantitative and that the acquisition and analysis of detailed temporal and spatial data will have to be a major focus of metabolomics specialists. While statistical, machine learning and chemometric methods for spectral analysis will likely continue to become more sophisticated, the long-term trend in metabolomics seems to be towards rapid/high throughput compound identification and quantification. These so-called targeted methods will require greater reliance on spectral libraries and spectral standards and will no-doubt lead to the appearance of organism-specific metabolite databases (just as there are organism-specific genome and proteome databases). This trend towards large-scale metabolite identification and quantification will likely encourage metabolomics specialists to adopt many of the analytical approaches commonly used in transcriptomics and proteomics, where transcript and protein levels are routinely quantified, compared and analyzed. Given the importance that classical bioinformatics has played in establishing genomics and proteomics as routine methods in modern biology, it is likely that continuing developments in computational metabolomics will be key to making the metabolomics a routine part of agricultural, nutritional, pharmaceutical and biomedical research.

Key Points

- Computational metabolomics requires a combination of both bioinformatics and cheminformatics.
- Metabolomics is a young field with many opportunities for computational development and informatics innovation.
- The areas of most active development over the past 2 years include metabolomics databases, data standards, data analysis tools and metabolic modeling.
- In the past year, a number of high-quality databases and LIMS have started to emerge and these are leading to improved data standards and improved data-sharing protocols for metabolomics researchers.
- Continuing developments in computational metabolomics will be key to making the field a routine part of agricultural, nutritional, pharmaceutical and biomedical research.

Funding

I would like to acknowledge Genome Canada, Genome Alberta and the Alberta Ingenuity Centre for Machine Learning (AICML) for their financial support.

References

1. German JB, Hammock BD, Watkins SM. Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics* 2005;**1**:3–9.
2. Kell DB. Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug Discov Today* 2006;**11**:1085–92.
3. Watkins SM, German JB. Metabolomics and biochemical profiling in drug discovery and development. *Curr Opin Mol Ther* 2002;**4**:224–8.
4. Schnackenberg LK, Beger RD. Monitoring the health to disease continuum with global metabolic profiling and systems biology. *Pharmacogenomics* 2006;**7**:1077–86.
5. Griffin JL, Bollard ME. Metabonomics: its potential as a tool in toxicology for safety assessment and data integration. *Curr Drug Metab* 2004;**5**:389–98.
6. Lindon JC, Holmes E, Nicholson JK. Metabonomics and its role in drug development and disease diagnosis. *Expert Rev Mol Diagn* 2004;**4**:189–99.
7. Nicholson JK, Connelly J, Lindon JC, *et al*. Metabonomics: a platform for studying drug toxicity and gene function. *Nat Rev Drug Discov* 2002;**1**:153–61.
8. Wishart DS, Querengesser LMM, Lefebvre BA, *et al*. Magnetic resonance diagnostics: a new technology for high-throughput clinical diagnostics. *Clin Chem* 2001;**47**: 1918–21.
9. Oostendorp M, Engelke UF, Willemsen MA, *et al*. Diagnosing inborn errors of lipid metabolism with proton nuclear magnetic resonance spectroscopy. *Clin Chem* 2006; **52**:1395–405.
10. Moolenaar SH, Engelke UF, Wevers RA. Proton nuclear magnetic resonance spectroscopy of body fluids in the field of inborn errors of metabolism. *Ann Clin Biochem* 2003;**40**: 16–24.

11. Bino RJ, Hall RD, Fiehn O, *et al*. Potential of metabolomics as a functional genomics tool. *Trends Plant Sci* 2004;**9**: 418–25.
12. Trujillo E, Davis C, Milner J. Nutrigenomics, proteomics, metabolomics, and the practice of dietetics. *J Am Diet Assoc* 2006;**106**:403–13.
13. Gibney MJ, Walsh M, Brennan L, *et al*. Metabolomics in human nutrition: opportunities and challenges. *Am J Clin Nutr* 2005;**82**:497–503.
14. Shulaev V. Metabolomics technology and bioinformatics. *Brief Bioinform* 2006;**7**:128–39.
15. Kanehisa M, Goto S, Hattori M, *et al*. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006;**34**(Database issue):D354–7.
16. Caspi R, Foerster H, Fulcher CA, *et al*. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* 2006;**34**(Database issue): D511–16.
17. Joshi-Tope G, Gillespie M, Vastrik I, *et al*. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 2005;**33**(Database issue):D428–32.
18. Crosby MA, Goodman JL, Strelets VB, *et al*. FlyBase: genomes by the dozen. *Nucleic Acids Res* 2007;**35**(Database issue):D486–91.
19. O'Donovan C, Martin MJ, Gattiker A, *et al*. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief Bioinform* 2002;**3**:275–84.
20. Wishart DS, Tzur D, Knox C, *et al*. HMDB: the Human Metabolome Database. *Nucleic Acids Res* 2007;**35**(Database issue):D521–6.
21. Smith CA, O'Maille G, Want EJ, *et al*. METLIN: a metabolite mass spectral database. *Ther Drug Monit* 2005;**27**: 747–51.
22. Seavey BR, Farr EA, Westler WM, *et al*. A relational database for sequence-specific protein NMR data. *J Biomol NMR* 1991;**1**:217–36.
23. Kopka J, Schauer N, Krueger S, *et al*. GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* 2005;**21**: 1635–8.
24. Duarte ND, Becker SA, Jamshidi N, *et al*. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Nat Acad Sci* 2007;**104**: 1777–82.
25. Spectral Database System (SDBS).Accessed from URL: http://www.aist.go.jp/RIODB/SDBS/cgi-bin/cre_index.cgi
26. Knox C, Shrivastava S, Stothard P, *et al*. BioSpider: a web server for automating metabolome annotations. *Pac Symp Biocomput* 2007;**12**:145–56.
27. Pearson H. Meet the human metabolome. *Nature* 2007; **446**:8.
28. Marshall E. Metabolic research: Canadian group claims 'unique' database. *Science* 2007;**314**:583–4.
29. Fiehn O, Kristal B, van Ommen B, *et al*. Establishing reporting standards for metabolomic and metabonomic studies: a call for participation. *OMICS* 2006;**10**: 158–63.
30. Castle AL, Fiehn O, Kaddurah-Daouk R, *et al*. Metabolomics standards workshop and the development of international standards for reporting metabolomics experimental results. *Brief Bioinform* 2006;**7**:159–65.
31. Brooksbank C, Quackenbush J. Data standards: a call to action. *OMICS* 2006;**10**:94–9.
32. Stromback L, Hall D, Lambrix P. A review of standards for data exchange within systems biology. *Proteomics* 2007;**7**: 857–67.
33. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. *Brief Bioinform* 2006;**7**:256–74.
34. Turner E, Bolton J. Required steps for the validation of a laboratory information management system. *Qual Assur* 2001;**9**:217–24.
35. Steinlechner M, Parson W. Automation and high throughput for a DNA database laboratory: development of a laboratory information management system. *Croat Med J* 2001;**42**:252–5.
36. White WL, Wagner CD, Hall JT, *et al*. Protein open-access liquid chromatography/mass spectrometry. *Rapid Commun Mass Spectrom* 2005;**19**:241–9.
37. Maurer M, Molidor R, Sturn A, *et al*. MARS: microarray analysis, retrieval, and storage system. *BMC Bioinformatics* 2005;**6**:101.
38. Goh CS, Lan N, Echols N, *et al*. SPINE 2: a system for collaborative structural proteomics within a federated database framework. *Nucleic Acids Res* 2003;**31**:2833–8.
39. Scholz M, Fiehn O. SetupX: A public study design database for metabolomic projects. *Pac Symp Biocomput* 2007;**12**: 169–80.
40. Markley JL, Anderson ME, Cui Q, *et al*. New bioinformatics resources for metabolomics. *Pac Symp Biocomput* 2007;**12**: 157–68.
41. Jenkins H, Hardy N, Beckmann M, *et al*. A proposed framework for the description of plant metabolomics experiments and their results. *Nat Biotechnol* 2004;**22**: 1601–06.
42. Zolnai Z, Lee PT, Li J, *et al*. Project management system for structural and functional proteomics: Sesame. *J Struct Funct Genomics* 2003;**4**:11–23.
43. Dunn WB, Bailey NJ, Johnson HE. Measuring the metabolome: current analytical technologies. *Analyst* 2005; **130**:606–25.
44. Trygg J, Holmes E, Lundstedt T. Chemometrics in metabonomics. *J Proteome Res* 2007;**6**:469–79.
45. Brindle JT, Nicholson JK, Schofield PM, *et al*. Application of chemometrics to 1H NMR spectroscopic data to investigate a relationship between human serum metabolic profiles and hypertension. *Analyst* 2003;**128**:32–6.
46. Weljie AM, Newton J, Mercier P, *et al*. Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Anal Chem* 2006;**78**:4430–42.
47. Lavine B, Workman JJ Jr. Chemometrics. *Anal Chem* 2004; **76**:3365–71.
48. Lindon JC, Holmes E, Nicholson JK. Metabonomics and its role in drug development and disease diagnosis. *Expert Rev Mol Diagn* 2004;**4**:189–99.
49. Holmes E, Nicholls AW, Lindon JC, *et al*. Chemometric models for toxicity classification based on NMR spectra of biofluids. *Chem Res Toxicol* 2000;**13**:471–8.
50. Smith IC, Baert R. Medical diagnosis by high resolution NMR of human specimens. *IUBMB Life* 2003;**55**:273–7.
51. Wilson ID, Plumb R, Granger J, *et al*. HPLC-MS-based methods for the study of metabonomics. *J Chromatogr B Analyt Technol Biomed Life Sci* 2005;**817**:67–76.

52. Fujiwara M, Ando I, Arifuku K. Multivariate analysis for 1H-NMR spectra of two hundred kinds of tea in the world. *Anal Sci* 2006;**22**:1307–14.
53. Gonzalez-Arjona D, Lopez-Perez G, Gonzalez-Gallero V, *et al*. Supervised pattern recognition procedures for discrimination of whiskeys from gas chromatography/mass spectrometry congener analysis. *J Agric Food Chem* 2006;**54**: 1982–9.
54. Gavaghan McKee CL, Wilson ID, *et al*. Metabolic phenotyping of nude and normal (Alpk:ApfCD, C57BL10J) mice. *J Proteome Res* 2006;**5**:378–84.
55. Liu L, Cozzolino D, Cynkar WU, *et al*. Geographic classification of Spanish and Australian tempranillo red wines by visible and near-infrared spectroscopy combined with multivariate analysis. *J Agric Food Chem* 2006;**54**: 6754–9.
56. Lutz U, Lutz RW, Lutz WK. Metabolic profiling of glucuronides in human urine by LC-MS/MS and partial least-squares discriminant analysis for classification and prediction of gender. *Anal Chem* 2006;**78**:4564–71.
57. Ruetschi U, Zetterberg H, Podust VN, *et al*. Identification of CSF biomarkers for frontotemporal dementia using SELDI-TOF. *Exp Neurol* 2005;**196**:273–81.
58. Wishart DS. Metabolomics: the principles and potential applications to transplantation. *Am J Transplant* 2005;**5**: 2814–20.
59. Robosky LC, Wells DF, Egnash LA, *et al*. Metabonomic identification of two distinct phenotypes in Sprague-Dawley (Crl:CD(SD)) rats. *Toxicol Sci* 2005;**87**:277–84.
60. Brown M, Dunn WB, Ellis DI, *et al*. A metabolome pipeline: from concept to data to knowledge. *Metabolomics* 2005;**1**:39–51.
61. Serkova NJ, Rose JC, Epperson LE, *et al*. Quantitative analysis of liver metabolites in three stages of the circannual hibernation cycle in 13-lined ground squirrels by NMR. *Physiol Genomics* 2007; (Epub ahead of print).
62. Serkova NJ, Zhang Y, Coatney JL, *et al*. Early detection of graft failure using the blood metabolic profile of a liver recipient transplantation. 2007;**83**:517–21.
63. Niwa T. Metabolic profiling with gas chromatography-mass spectrometry and its application to clinical medicine. *J Chromatogr* 1986;**379**:313–45.
64. la Marca G, Casetta B, Malvagia S, *et al*. Implementing tandem mass spectrometry as a routine tool for characterizing the complete purine and pyrimidine metabolic profile in urine samples. *J Mass Spectrom* 2006;**41**:1442–52.
65. Jiang H, Somogyi A, Timmermann BN, *et al*. Instrument dependence of electrospray ionization and tandem mass spectrometric fragmentation of the gingerols. *Rapid Commun Mass Spectrom* 2006;**20**:3089–100.
66. Ott KH, Aranibar N, Singh B, Stockton GW. Metabonomics classifies pathways affected by bioactive compounds. Artificial neural network classification of NMR spectra of plant extracts. *Phytochemistry* 2003;**62**: 971–85.
67. Alves R, Antunes F, Salvador A. Tools for kinetic modeling of biochemical networks. *Nat Biotechnol* 2006;**24**(6):667–72.
68. Materi W, Wishart DS. Computational systems biology in drug discovery and development: methods and applications. *Drug Discov Today* 2007;**12**:295–303.
69. Mendes P. GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. *Comput Appl Biosci* 1993;**9**:563–71.
70. Kitano H, Funahashi A, Matsuoka Y, *et al*. Using process diagrams for the graphical representation of biological networks. *Nat Biotechnol* 2005;**23**:961–6.
71. Sauro HM. SCAMP: a general-purpose simulator and metabolic control analysis program. *Comput Appl Biosci* 1993;**9**:441–50.
72. Shapiro BE, Levchenko A, Meyerowitz EM, *et al*. Cellerator: extending a computer algebra system to include biochemical arrows for signal transduction simulations. *Bioinformatics* 2003;**19**:677–8.
73. Demir O, Aksan Kurnaz I. An integrated model of glucose and galactose metabolism regulated by the GAL genetic switch. *Comput Biol Chem* 2006;**30**:179–92.
74. Mendoza-Cozatl DG, Moreno-Sanchez R. Control of glutathione and phytochelatin synthesis under cadmium stress. Pathway modeling for plants. *J Theor Biol* 2006;**238**: 919–36.
75. Gagneur J, Casari G. From molecular networks to qualitative cell behavior. *FEBS Lett* 2005;**579**:1867–71.
76. Joyce AR, Palsson BO. Toward whole cell modeling and simulation: comprehensive functional genomics through the constraint-based approach. *Prog Drug Res* 2007;**64**:267–309.
77. Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. *Curr Opin Biotechnol* 2003;**14**:491–6.
78. Lee JM, Gianchandani EP, Papin JA. Flux balance analysis in the era of metabolomics. *Brief Bioinform* 2006;**7**:140–50.
79. Marx A, Eikmanns BJ, Sahm H, *et al*. Response of the central metabolism in *Corynebacterium glutamicum* to the use of an NADH-dependent glutamate dehydrogenase. *Metab Eng* 1999;**1**:35–48.
80. Oliveira AP, Nielsen J, Forster J. Modeling *Lactococcus lactis* using a genome-scale flux model. *BMC Microbiol* 2005;**5**:39.
81. Price ND, Thiele I, Palsson BO. Candidate states of *Helicobacter pylori*'s genome-scale metabolic network upon application of 'loop law' thermodynamic constraints. *Biophys J* 2006;**90**:3919–28.
82. Borodina I, Krabben P, Nielsen J. Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. *Genome Res* 2005;**15**: 820–9.
83. Edwards JS, Palsson BO. Metabolic flux balance analysis and the in silico analysis of *Escherichia coli* K-12 gene deletions. *BMC Bioinformatics* 2000;**1**:1.
84. Segre D, Deluna A, Church GM, *et al*. Modular epistasis in yeast metabolism. *Nat Genet*. 2005;**37**:77–83.
85. Jin YS, Jeffries TW. Stoichiometric network constraints on xylose metabolism by recombinant *Saccharomyces cerevisiae*. *Metab Eng* 2004;**6**:229–38.
86. Durmus Tekir S, Cakir T, Ulgen KO. Analysis of enzymopathies in the human red blood cells by constraint-based stoichiometric modeling approaches. *Comput Biol Chem* 2006;**30**:327–38.
87. Luo RY, Liao S, Tao GY, *et al*. Dynamic analysis of optimality in myocardial energy metabolism under normal and ischemic conditions. *Mol Syst Biol* 2006;**2**:2006.0031.
88. Kell DB. Metabolomics, machine learning and modelling: towards an understanding of the language of cells. *Biochem Soc Trans* 2005;**33**:520–4.