Design and Development of Novel Metabolomic Databases and Tools

by

Timothy Jewison

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

Abstract

Metabolomics involves the high throughput characterization of small molecules or metabolites in cells, tissues and organisms. To interpret, store and exchange metabolomic data it is necessary to have comprehensive, electronically accessible databases that can be used to handle both the experimental data and the associated biological and chemical information needed to identify, quantify and interpret the meaning of those metabolites. As the field of metabolomics matures the need for improved databases is growing rapidly. In particular, there is a serious shortage of organism-specific metabolomic databases and a significant bottleneck with regard to the breadth and depth of known metabolic pathways needed to interpret metabolomic data. Consequently the main objective of my thesis project was to develop novel software and innovative databases to address these two metabolomics bottlenecks. In particular, I focused on the development of (i) species specific compound databases, (ii) the creation of graphical pathway databases and (iii) the design and implementation of innovative pathway visualization techniques and tools. This thesis describes the design and implementation of the Yeast Metabolome Database (YMDB) (`http://ymdb.ca`) an example of an organism-specific metabolomic database, and the Small Molecule Pathway Database (SMPDB) (`http://smpdb.ca`) an example of a comprehensive graphical pathway database. It also describes the PathWhiz system, a novel web server for the creation and visualization of biological pathways.

Preface

Chapter 2 was previously published in Nucleic Acids Research as: Jewison T, Knox C, Neveu V, Djoumbou Y, Guo AC, Lee J, Liu P, Mandal R, Krishnamurthy R, Sinelnikov I, Wilson M, Wishart DS. (2012) YMDB: the Yeast Metabolome Database. Nucleic Acids Res. 40(Database issue):D815-20. As the first author I was responsible for designing and assembling the database. Specifically I researched, found, entered and validated most of the data, coordinated the data entry activities with the annotation staff, designed and built the web interface and developed software to assist with the data mining, formatting and deposition of the data. I was also responsible for preparing the first draft of the manuscript, for producing the tables and figures and for providing feedback and edits through all later drafts of the paper.

Chapter 3 was previously published in Nucleic Acids Research as: Jewison T, Su Y, Disfany FM, Liang Y, Knox C, Maciejewski A, Poelzer J, Huynh J, Zhou Y, Arndt D, Djoumbou Y, Liu Y, Deng L, Guo AC, Han B, Pon A, Wilson M, Rafatnia S, Liu P, Wishart DS. (2014) SMPDB 2.0: Big Improvements to the Small Molecule Pathway Database. Nucleic Acids Res. 42(1):D478-84. As the first author I was responsible for designing and assembling the database and the software necessary to submit data into the database. Specifically I helped illustrate and validate most of the pathway and ancillary data in the database, coordinated the data entry activities with the annotation staff, designed and built the web interface and developed software to assist with the pathway illustration, formatting and deposition of the pathway data. I was also responsible for preparing the first draft of the manuscript, for producing the tables and figures and for providing feedback and edits through all later drafts of the paper.

Chapter 4 will be submitted to Nucleic Acids Research as: Jewison T, Su Y, Disfany FM, Liang Y, Knox C, Maciejewski A, Wilson M and Wishart DS. PathWhiz: a Web Server for Pathway Generation and Visualization. As the first author I was responsible for designing and assembling the web server. Specifically I designed and built the web interface and developed all of the software needed to draw, enter, visualize and save the pathway diagrams. I was responsible for testing the software and optimizing the code and its performance based on feedback from my co-authors. I was also responsible for preparing the first draft of the chapter, for producing the tables and figures and for providing feedback and edits through all later drafts of the chapter.

Acknowledgements

I would like the thank my supervisor Dr. David S. Wishart who supported and encouraged me in my education. He provided the guidance and structure necessary for me to excel in my graduate studies. I am very grateful for the opportunity he provided me to work with a diverse group of scientists, on many interesting and challenging projects.

I would also like to thank the members of the Wishart lab for their hard work and patience while assisting me. In particular, I would like to thank lab veteran Craig Knox. He was critical in providing me with the background knowledge needed to succeed in the development of my projects. He also provided day-to-day feedback and suggestions throughout my activities.

Overall, there were many challenges in completing this degree, and I never could have done it without the support and assistance of my colleagues, friends, and family.

Table of Contents

List of Tables

List of Figures

List of Terms and Abbreviations

CSV	Comma Seperated Value
DNA	Deoxyribonucleic Acid
GC-MS	Gas-Chromatography combined with Mass Spectrometry
GUI	Graphical User Interface
IEM	Inborn Errors of Metabolism
NMR	Nuclear Magnetic Resonance
LC-MS	Liquid-Chromatography combined with Mass Spectrometry
MS	Mass Spectrometry
MS/MS	Tandom Mass Spectrometry
PNG	Portable Network Graphics
RNA	Ribonucleic Acid
SBML	Systems Biology Markup Language
SMILES	Simiplified Molecular Input Line Entry Specification
SOP	Standard Operating Procedure
SVG	Scalable Vector Graphics
URL	Uniform Resource Locator
XML	Extensible Markup Language

Chapter 1

Introduction

1.1 The Metabolome

Metabolites are the small molecules consumed or generated through a combination of metabolic and catabolic processes within living systems. Metabolites include low molecular weight (< 1500 Daltons) molecules such as amino acids, nucleotides, fatty acids, lipids, vitamins and minerals. Generally large biopolymers such as proteins, deoxyribonucleic acids (DNA), ribonucleic acids (RNA) and complex sugars like starch are not considered to be small molecules, nor are they considered metabolites. The set of all metabolites associated with a particular biological system, including cells, tissues, biological extracts and whole organisms, is called the metabolome [19]. The metabolome can be further partitioned into two broad categories: the endogenous metabolome and exogenous metabolome. Endogenous metabolites are those generated by an organism's natural metabolic or catabolic processes while exogenous metabolites are metabolites that come from external or man-made sources including food, toxins, drugs and the environment. Metabolomics is a field of omics science that focuses on the characterization, detection and quantification of metabolites and metabolomes. Metabolomics is a relatively new field when compared to its more mature cousins of genomics, transcriptomics and proteomics.

Metabolomics differs from other omics fields in that the analytical techniques used to quantify and qualify metabolites are often specific to certain types or classes of compounds. In contrast, for genomics, transcriptomics and proteomics, the de-

tection techniques are generally nonspecific or universal. For instance, in metabolomics NMR (Nuclear Magnetic Resonance) is particularly suited to detecting higher concentrations of water-soluble metabolites (such as alcohols and sugars), while mass spectrometry (MS) is best for detecting low concentrations of hydrophobic metabolites (such as lipids or organic acids) while gas-chromatography combined with mass spectrometry (GC-MS) is best for detecting volatile or gas-phase metabolites. Thus there are many different approaches to detecting metabolites and each varies with its level of sensitivity, its mode of detection or the compound classes that it can detect. The commonality between most metabolomics techniques is the production of a complex spectrum or a chromatogram that can be used to identify and possibly quantify a metabolite or a set of metabolites. When multiple analytical techniques are combined (NMR + MS + GC-MS) the entire metabolome, or at least a good portion of it, can be characterized.

A generalized metabolomics workflow is shown in Figure 1.1. This first step of the workflow is to define the problem that needs to be solved. In many clinical studies this might involve performing a case-control analysis where samples from healthy individuals are collected and compared against samples from individuals affected by a well-defined disorder. Once the problem or subject group has been identified, the second phase is sample collection. The sample collection phase involves collecting a physical sample from each study subject, this can be tissue samples, biofluid samples (blood, urine, etc.) or other biological materials. In the third step the biosample must be prepared for spectral collection. How the sample is prepared in this phase will depend on the types metabolites that need to be detected as well the analytical techniques being used to measure the metabolites. In almost all cases a spectrum or a chromatogram with hundreds to thousands of peaks (with each peak corresponding to a different compound or metabolite feature) is generated. Once these spectra have been collected and processed they can then be used to identify and/or quantify the metabolites in the biosample. This step is both the most difficult and time-consuming process in metabolomics as it typically involves a combination of tedious manual and computationally intense efforts. Most compound identification techniques involve spectral deconvolution, spectral fitting and

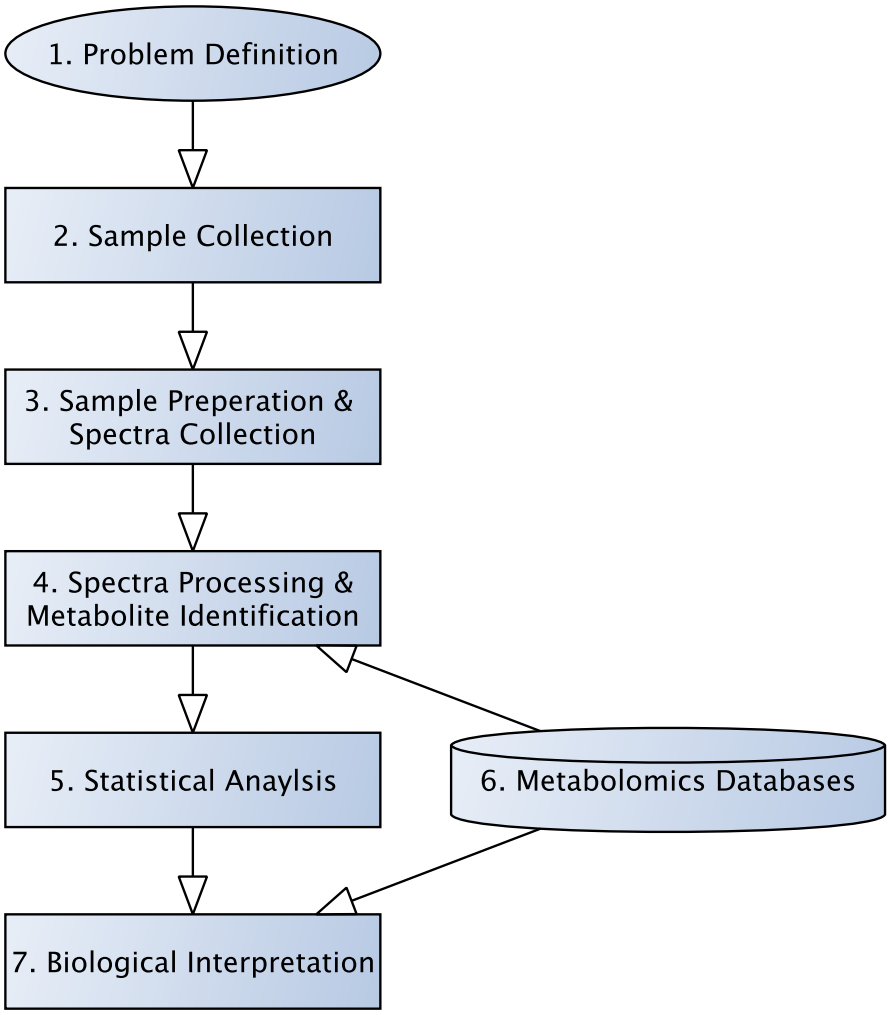

Figure 1.1: A typical metabolomics workflow

spectral comparison to a reference database of pure compounds. This process can take hours or days for each spectrum and its success is critically dependent on the quality and comprehensiveness of the database or databases being used. Once all metabolites in the samples of interest have been identified and quantified, statistical analysis can be done to evaluate to differences between samples and controls. This final step typically makes use of multivariate statistical techniques to identify statistically significant metabolic differences. This information can then be further interpreted through the use of metabolic or signaling pathway analysis. It is this final biological interpretation phase which also requires considerable manual effort.

Indeed, the success or failure of many metabolomic and pathway analyses often depends on the quality and comprehensiveness of the pathway databases being used.

The importance of referential databases both in spectral analysis (step #3) and in interpretation (step #6) cannot be emphasized enough. Without these databases or without access to improved databases, most of the key aspects of modern metabolomics could not be performed. The role and the importance of metabolomic and pathway databases will be further discussed in Section 1.3.

1.2 Metabolomics Applications

Metabolites represent the end product of many cell-level processes, including cellular metabolism and catabolism, as well as numerous physiological process such as breathing, ingestion, digestion, and elimination. The metabolites produced by an organism are a result of both its genetics (plants produce different metabolites than animals) and its environment (we are what we eat). Thus an organism's metabolome closely reflects its phenotypic state (or its Phenome). Indeed, among all the omics fields, metabolomics is truly unique in its ability to measure gene-environment interactions and to provide a clear phenotypic readout. This capacity makes metabolomics particularly useful in the characterization or identification of both disease and phenotypic biomarkers [64]. Metabolomics is seeing widespread applications in the fields of clinical medicine, veterinary medicine, drug testing and assessment, food and nutritional chemistry, dietary assessment, food and beverage production, toxicology and environmental monitoring [69, 50, 85, 84, 68, 34, 1, 24, 4].

One of the first applications of metabolomics was in the identification and characterization of human metabolic disorders [82]. Metabolic disorders, or inborn errors of metabolism (IEMs), are fundamentally the result of deleterious mutations leading to perturbations and dysregulation of key metabolic pathways. IEMs are relatively rare and are often characteristically unique to an individual, making them difficult to diagnose and treat [23]. Early on it was recognized that metabolomics could offer clinical chemists the opportunity to detect and characterize clinically difficult metabolic perturbations or dysregulation because it measures such a wide

range of chemical classes. As a result, metabolomics has been used to discover and help treat a number of IEMs including phenylketonuria [15, 38], medium-chain acyl-coenzyme A dehydrogenase deficiency[15, 38], methylmalonic aciduria [38, 82], maple syrup urine disease [38], and propionic acidemia [82]. Based on these early successes, many are expecting that clinical metabolomic analyses combined with next-generation DNA sequencing will soon lead to a new era of personalized medicine that will benefit many individuals suffering from IEMs and other hard-to-diagnose metabolic disorders [3, 22, 76].

Another common application of metabolomics has been in the wine, beer and spirits industries. The fermentation of alcoholic beverages is critically dependent on yeast metabolism. When yeast are grown in anaerobic conditions with grape must, grains or other plant products they not only produce ethanol but also a variety of other aromatic or flavour-enhancing compounds from the secondary metabolites found in these plants. The process of producing high-value alcoholic beverages with complex flavours and aromas is function of fermantation substrates, the particular metabolic capability or strain of the yeast used and the physical environment used for production. However, until recently it was very difficult to identify the specific flavour compounds or to associate specific fermentation substrates, processes and pathways that affected the production of these compounds. Metabolomic techniques have been used to clarify many of these issues and to help characterize and classify wines [34], beers [37] and scotch whiskeys [31]. In this way metabolomics is helping food and beverage researchers better understand and potentially improve the fermentation process.

The successful application of metabolomics to such diverse disciplines as IEM detection and fermented beverage testing have depended critically on the availability of specialized metabolomic databases. These metabolomic databases may include spectral databases (for compound ID), pathway databases (for biological interpretation), compound databases (for improved chemical understanding) and reaction databases (for biochemical interpretation). The next section discusses metabolomic databases in more detail and provides a more comprehensive overview of what kinds of databases already exist and which kinds are needed.

1.3 Metabolomic Databases

There are three types of metabolomic databases: 1) chemical databases, 2) spectral databases, and 3) pathway databases [61]. Each particular database category fills a particular need in analyzing and interpreting metabolomic data. I will review each of these database categories separately and provide a few examples of each.

1.3.1 Chemical Databases

Chemical databases are primarily used to assist in the identification of metabolites. This is done by way of matching an observed chemical structure, a calculated atomic composition, a measured mass or some observed physical properties of an unidentified compound to something already in the database. Historically, chemical databases consisted of collections of 1000's or 10's of thousands of chemical structures with detailed information about compound names, structures, molecular weight, predicted solubility, estimated pKa, etc. along with reference experimental properties like NMR and MS spectra, chromatographic retention times, and MS/MS fragmentation patterns [59, 25]. Until recently most chemical compound databases could only be found in books and journals. However, with the rise of the internet it has now become practical to create web accessible databases that can store far more chemical compound data than can be found in any book or journal. Furthermore, these web-based systems are proving to be far more accessible and much easier to query than any hard-cover reference manual. These web-based databases can centralize and aggregate data, thereby allowing for the rapid querying and retrieval of data, a process that is extremely time consuming when done through manual literature searching.

Chemical databases can be further divided into three subcategories: 1) general chemical repositories, 2) biological specific databases and 3) organism specific databases. General chemical repositories include the NIST Chemistry WebBook [55], PubChem [78] and Chemspider [62]. Chemical repositories serve as general repositories of chemical information and contain many entries that are not metabolites and not natural products. They may include synthetic or man-made pesticides

or herbicides, petroleum products, plastics, toxins, experimental chemicals, synthetic chemistry reagents and organic solvents. Because of their broad scope, most chemical databases contain millions of compounds. However, less than 1% of these compounds are biologically relevant. Biological-specific databases include ChEBI [13, 30, 12] and the KEGG LIGAND database [44]. Biological-specific databases focus on archiving compounds that are natural products excluding synthetic or man-made compounds or those that play no biological role. These databases typically include compounds from plants, animals and bacteria and do not typically distinguish where the compounds originated or whether they are unique to a specific organism or a specific class of organism. Most biological-specific databases contain 10,000+ compounds. Organism specific databases include the Human Metabolome Database (HMDB) [86, 87, 89], the Yeast Metabolome Database (YMDB) [40] and the Escherichia coli Metabolome Database (ECMDB) [28]. Organism specific databases are similar to biological specific databases but only include compounds that have been confirmed to be in or produced by that specific organism.

Chemical databases can also be supplemented by chemical libraries. Chemical libraries are collections of authentic samples of pure compounds that have either been synthesized or purified by chemists. Chemical libraries are frequently used in metabolomic studies to confirm compounds through so-called spike-in experiments. Spike-in experiments or guess-and-check experiments involve dropping small amounts of the authentic material (or isotopically labeled forms of the material) into a biological sample and seeing if the signal corresponding to the spiked-in material matches the signal for the unknown compound. However, chemical libraries are expensive to build and challenging to maintain. Consequently, most metabolomics researchers depend on chemical databases to help with compound identification.

1.3.2 Spectral Databases

Spectral databases are used to aid in the identification of metabolites by providing reference NMR, MS or IR spectra of purified compounds along with other relevant experimental properties. Spectral databases can be used to identify unknown

metabolites, to confirm the identity of a known metabolite or to narrow down possible candidates. Spectral databases including the NIST/EPA/NlH Mass Spectral Database, the Golm Metabolomics Database (GMD) [53] and Massbank [35] provide reference GC-MS, LC-MS and MS/MS spectra for a variety of MS based techniques. References for small molecule NMR spectra can be found in the Biological Magnetic Resonance DataBank (BRMB) [74], NMRShiftDB [70] and the Spectral Database for Organic Compounds (SDBS) (`http://sdbs.riodb.aist.go.jp`). Most spectral databases contain reference spectra for not only metabolites but also synthetic, industrial and other non-natural compounds that would not likely be biologically relevant. Searching these databases to identify an unknown compound can be a challenge due to the potential generation of false positives associated with compounds that are not biologically relevant. This recurring problem has led to the development of chemical or spectral databases that are biological or organism-specific. These kinds of biologically-oriented spectral databases provide search capabilities and reference spectra that are only relevant in a biological or organism-specific context. Chemical/spectral databases such as HMDB, YMDB, and ECMDB are examples of databases that provide organism-specific spectral information.

1.3.3 Pathway Databases

Pathway databases are a very distinct category of metabolomic database. They are more applicable to analyzing and interpreting metabolomic results once a metabolite or set of metabolites has been identified. Pathway databases are particularly useful for determining the contextual role of metabolites in a biological system and for interpreting the changes in metabolism associated with these metabolites. Pathways explain biological processes in a step by step fashion that lead to a particular outcome. Most pathway databases provide visual descriptions of metabolic or signaling pathways that can aid in the interpretation of biological data or the understanding of biological phenotypes. Commercial pathway databases including those offered by Sigma-Aldrich (`http://www.sigmaaldrich.com`), Millipore (`http://www.millipore.com`) and Bio-

carta (http://www.biocarta.com) focus on integrating product sales with pathway visualizations. Consequently these commercial databases are always visually impressive and hyperlinked to plenty of commercial products. However they are generally less informative or less comprehensive than publicly-accessible databases. Most commercial pathway databases generally focus on non-metabolic (i.e. protein or gene-signalling) pathways. Public databases including The Small Molecule Pathway Database (SMPDB) [20], KEGG [44], MetaCyc [46], Wikipathways [48], and Reactome [57] provide a rich variety of pathway visualizations for a wide range of organisms. Furthermore, these pathway databases are generally much more focused on metabolism. Obviously there are many more public pathway databases than those listed here. A more complete listing of available pathway databases can be found at the PathwayCommons [8].

1.4 Database Curation Challenges

Assembling and maintaining a complex database typically takes a skilled team of annotators and programmers. One of the biggest difficulties with creating and maintaining chemically oriented databases is dealing with the unique identification of compounds. The best and most accurate way to uniquely identify a compound is by its chemical structure. Unfortunately the chemical structure or stereochemistry of a compound is not always fully known nor is it necessarily complete. As a result, most compounds are referenced only by name. To complicate matters further it is common to use names of compounds that represent mixtures or that describe a general, non-unique, structure. Currently there is no chemical naming convention that is independent of the chemical structure. Chemical naming is also influenced by regional, commercial and historical events that can inadvertently lead to many different synonyms being developed for any given compound. A typical example of this is seen with amino acids. Alanine is the general name of L-Alanine and D-Alanine which are stereoisomers (mirror images) of each other. When discussing L-Alanine within the context of human biology it is generally acceptable to use the term Alanine because humans can only produce L-Alanine naturally. However,

bacteria can also produce D-Alanine and so when one is looking at bacterial products coming from humans, it is possible for the naming situation to become quite confusing.

The Chemical Abstract Service (CAS) Registry attempted to solve this problem by assigning all compounds (found in scientific literature since the early 1900s) with a unique CAS identifier or CAS number. CAS numbers are considered the world standard for identifying compounds, though their use is far from standardized. For instance, most metabolomics papers are published without reference CAS numbers being associated with their compound lists. This is likely because the CAS registry is a commercial service and there is no global governing body enforcing their use. With the creation of many different types of chemical based databases with their own unique identifiers the identification of compounds by name has become a non trivial task for all but the commonest or best characterized chemical compounds.

Another major challenge for constructing and maintaining metabolomic (as well as other omic) databases is the extraction and collection of information. This often requires significant resources and a long or lengthy manual (curatorial) effort. A common solution to this problem is allowing the public to deposit their data directly in a database. The strategy has been adopted by both PubChem and MassBank. One problem with this solution is a cultural one — public deposition relies on the desire of individual researchers to make their data publicly available. This problem does not exist in the field of genomics and proteomics where sequences generally have to be submitted to a database before a paper is accepted for publication. However, in the fields of chemistry, analytical chemistry, natural product chemistry and metabolomics, no such publication obligation or cultural desire exists. Consequently most chemical data still resides in books and journals and the curation or extraction of this data still requires significant manual transcription and long periods of careful reading and manual entry.

The problem of data extraction is a complex one. One particular challenge with chemical data extraction is cross referencing compound names (which are the typical identifier used in a journal article) with a discrete chemical structure. This problem extends from the long history of chemistry (especially biological chemistry)

where chemical naming was not a systematic or well-defined process. Chemical names do not always uniquely identify a real compound. This makes it difficult to discern the identity of a compound and it also makes it difficult to ensure that any accessory data that is being ascribed to the compound is correct. Once again, the only solution to this problem lies in the use of expert human curators with considerable domain knowledge.

While the challenges associated with building and maintaining metabolomic (or chemical) databases are significant, they are potentially even greater with the development and maintenance of pathway databases. This is because pathways are essentially an artists rendering of a biological process. They are not discrete, well defined entities like chemical structures, formulas or reactions. Indeed, most biological pathways are only incomplete (and often incorrect) simplifications of what we think is going on inside a cell. Furthermore, the start and end or extent of a biological pathway is often arbitrarily defined by the energy and interest of the pathway artist. The fact that pathways and pathway diagrams are more art than science has meant that in the world of biological pathways and pathway databases there is relatively little standardization in how pathways are represented or how they are exchanged. As a result, the world of pathway databases and the exchange of pathway data is very similar to the Wild West of the 1850s.

1.5 Data Exchange

Data exchange and data dissemination is an essential part of the scientific process. While publication and public presentations are the most common ways of disseminating and assessing processed scientific data, there is a growing realization within the scientific community that papers, alone, cannot capture or present all of the scientific information generated by many modern omics experiments. Indeed, the quantity, diversity and breadth of data being generated in many omics studies including metabolomics, is such that only a tiny fraction of the data being generated is being made publicly available through papers. One way to address this data deficit is to deposit or disseminate experimental omics data through publicly ac-

cessible databases and to make these data available in a consistent, complete and standardized format. As a result, data exchange along with standardized data exchange formats and standardized ontologies are being developed and tested through a variety international efforts. Providing data in an exchangeable format allows for the assembly of derived data sets from multiple sources. Furthermore it allows the facile import of new data into existing datasets or databases. Given their many advantages and benefits, most life science fields are developing data exchange formats and controlled vocabularies to identify common concepts and concept relationships (i.e. ontologies). Ontologies can be used to annotate and describe data as well as its relationships to concepts. Ontologies, when combined with a general data exchange formats like XML form an excellent mechanism for storing, sharing and working with experimental data and deriving knowledge from that data.

Over the past decade a number of ontologies and data exchange standards have been designed and proposed for the exchange of chemical, spectral and pathway data. The Chemical Markup Language (CML) [60] is an XML based format used to describe both small molecules and macromolecules in a consistent and extensible manner. MzML [56] is an example of another XML based format designed for describing mass spectral data for both proteomics and metabolomics applications. Currently a new data exchange format called nmrML (`http://nmrml.org`) is being developed to describe NMR spectra specifically for metabolomics experiments. Another data exchange format called BioPAX [14] is a RDF/XML format that can represent biological processes and biological pathway data. While each of these data exchange formats has been carefully designed and extensively tested, there continues to be a problem with their widespread adoption and relatively slow uptake especially within the metabolomics community. Two of the reasons for this slow uptake appear to be 1) the lack of useful data sets or popular databases using these formats and 2) the limited number of tools that allow scientists to easily transfer their data into these formats.

1.6 Thesis Motivation and Problem Statement

In the previous 5 sections I have highlighted some of the key bottlenecks and central, unsolved challenges in the field of metabolomics. Many of these problems are centered around the lack of appropriate data or the lack of data in appropriate formats. One clear bottleneck in metabolomics lies in metabolite identification. This is compounded by the lack of organism-specific databases to aid in the identification or characterization of metabolites for important model organisms. A second bottleneck for metabolomics is the lack of comprehensive pathway databases with robust querying capabilities to permit facile biological interpretation of metabolomic data. A third bottleneck or unsolved challenge in metabolomics is the lack of popular or comprehensive databases that make use of standard data exchange formats. This is preventing many metabolomics researchers from accessing or using a number of key metabolomic data resources. Finally, a fourth unsolved challenge is the lack of software tools to permit the facile generation of biological pathways in standard or exchangeable pathway formats.

Based on these four problems or challenges I have developed a number of potential solutions. In particular, to address the first bottleneck (compound ID) I have developed an organism-specific metabolomic database called the yeast metabolome database (YMDB). This web-accessible database contains the largest compilation of yeast-specific metabolites and yeast biochemical reactions ever compiled. It also contains an extensive set of reference spectra to aid in yeast metabolite identification. In addition to requiring significant manual curation efforts, this database also led to the development of a number of novel software tools to aid in the compilation, formatting and querying of the yeast metabolite data.

To address the second bottleneck (better pathway databases) I have developed a substantially improved and updated version of the Small Molecule Pathway Database (SMPDB). By expanding the database by almost a factor of two (to cover more than 600 human metabolic pathways) and by substantially enhancing the querying and visualization capabilities within the database, I believe I have been able to make pathway querying and pathway interpretation far easier and far more robust.

To address the third bottleneck (lack of data in standard exchange formats) I have developed several databases (YMDB, ECMDB and SMPDB) that contain extensive quantities of data in community-approved standard exchange formats. This includes BioPAX/SBML for the pathway data contained in SMPDB, mzML and nmrML for the spectral data contained in YMDB ad ECMDB and InCHI for the chemical structure data contained in SMPDB, YMDB and ECMDB.

Finally, to address the fourth bottleneck or challenge (lack of tools for standardized pathway generation) I have developed a software package called PathWhiz. This interactive, user-friendly, web-based tool was used to generate the 600+ high quality pathways in SMPDB and to format each pathway into a standard BioPAX format. This unique system permits the rapid, facile generation of richly illustrated, colorful pathways with detailed chemical structures and extensive hyperlinks to external databases (UniProt and HMDB)

1.7 Thesis Outline

Because much of the work I have completed for this thesis has either been published or is about to be submitted for publication, I have decided to prepare a paper-based thesis. The thesis is organized as follows: Chapter 1 provides a general introduction to metabolomics. It also describes a number of bioinformatics or data challenges facing metabolomics researchers. Chapter 2 describes the Yeast Metabolome Database (YMDB) and is largely derived from my paper describing the database that was published in the 2012 database issue of Nucleic Acids Research. Chapter 3 describes the second version of the Small Molecule Pathway Database (SMPDB) which was published in the 2014 database issue of Nucleic Acids Research. Chapter 4 describes the software tool or webserver called PathWhiz. This chapter has been formatted to be consistent with the formatting requirements for the Web Server issue of Nucleic Acids Research. Chapter 5 provides some general conclusions and possible future work.

Chapter 2

YMDB: the Yeast Metabolome Database

2.1 Introduction

Metabolomics is a field of 'omics' research that is primarily focused on the identification and characterization of small molecule metabolites in cells, organs and organisms [80]. Together with genomics, transcriptomics and proteomics these four 'omics' disciplines form the cornerstones to systems biology. However, relative to its more mature 'omics' cousins, metabolomics still lags far behind in developing or formalizing its software and database infrastructure [83]. This is because the needs of metabolomics researchers span a very diverse range of scientific disciplines including organic chemistry, analytical chemistry, biochemistry, molecular biology and systems biology. In other words, metabolomics requires a tight blending of the tools found in both bioinformatics and cheminformatics. To address these informatics challenges, we (and others) have been steadily developing a set of comprehensive and open access tools to lay a more solid software/database foundation for metabolomics [83, 90, 92]. In particular, our group has developed several widely used organism- or discipline-specific databases including the Human Metabolome Database (HMDB) [87], DrugBank [88], the CyberCell database (CCDB) [72], the Toxin/Toxin-Target database (T3DB) [54] and the Small Molecule Pathway Database (SMPDB) [20]. HMDB, T3DB, DrugBank and SMPDB were specifically developed to address the metabolomics, toxicology, pharmacology and systems biology associated with humans (i.e. *Homo sapiens*), whereas CCDB was specifically

developed to address the metabolomics and systems biology needs for *Escherichia coli*.

We believe that the establishment and maintenance of organism-specific metabolomics databases is absolutely critical to the field of metabolomics as each organism has a unique and chemically distinct metabolome. The 'naïve' identification of metabolites, by simple mass matching for instance, without regard to their origin (organism or man-made) frequently leads to spurious, humorous or meaningless compound identifications [66]. Therefore, as part of our ongoing effort to create species-specific metabolomic resources for other model organisms we have now turned our attention to yeast, or more specifically, *Saccharomyces cerevisiae*.

The metabolic byproducts of *S. cerevisiae* fermentation are particularly interesting from both a biochemical and an industrial point of view. Indeed, *S. cerevisiae* (and its various strains) is perhaps the world's most important microbial biofactory, playing a key role in industrial chemical or biofuel production (ethanol), in the baking industry, as well as in beer, wine and spirit production. Together, these yeast-based industries are worth more than one trillion dollars per year to the global economy [39]. As a model organism for molecular biologists, *S. cerevisiae* is certainly the most intensively studied microbe and perhaps the most well understood living thing on earth. Being one of the first organisms to be fully sequenced [26] and being particularly amenable to unique and powerful genetic manipulations [10, 9] the sequence, function and interacting partner(s) of every gene/protein in *S. cerevisiae* is now almost completely known. This knowledge is contained in a number of excellent yeast-specific resources including SGD [18], YPD [33], CYGD [27] and FunSpec [65]. This remarkably detailed molecular knowledge has also made *S. cerevisiae* a favorite model organism for systems biologists, leading to the development of some very useful resources aimed at modeling or describing yeast pathways and metabolism including YeastNet [32], MetaCyc [5], KEGG [43] and Reactome [42]. Each of these excellent databases contains valuable information on primary yeast metabolic reactions, pathways and primary yeast metabolites.

Unfortunately, none of these systems biology databases contains information on the secondary metabolites of yeast fermentation (those compounds that give

wine, beer and certain cheeses or breads their flavor or aroma), yeast-specific lipids, yeast volatiles or yeast-specific ions. These actually represent hundreds of industrially and biochemically important compounds. Furthermore, none of today's current set of yeast systems biology databases provides detailed metabolite descriptions, intra- or extra- cellular concentrations, growth conditions, physico-chemical properties, subcellular locations, reference Nuclear Magnetic Resonance (NMR) or Mass Spectrometry (MS) spectra or other parameters that might typically be needed by researchers interested in yeast metabolism or yeast fermentation. For metabolomics researchers, as well as industrial chemists working with yeast byproducts, these kinds of data need to be readily available, experimentally validated, fully referenced, easily searched and readily interpreted. Furthermore, they need to cover as much of the yeast metabolome as possible. In an effort to address these shortcomings with existing yeast systems biology databases and to create a database specifically targeting the needs of yeast metabolomics, we have developed the Yeast Metabolome Database (YMDB).

2.2 Database Description

The YMDB is a combined bioinformatics-cheminformatics database with a strong focus on quantitative, analytic or molecular-scale information about yeast metabolites and their associated properties, pathways, functions, sources, enzymes or transporters. The YMDB builds upon the rich data sets already assembled by such resources as YeastNet 4.0 [32], MetaCyc [5], KEGG [43], UniProt [2], ChEBI [13] and HMDB [87]. But it also brings in a large body of independently collected literature data, as well as a significant quantity of experimental data, including NMR spectra, MS spectra and validated metabolite concentrations, to compliment this electronic or literature-derived data.

The diversity of data types, the quantity of experimental data and the required breadth of domain knowledge made the assembly of the YMDB both difficult and time-consuming. To compile, confirm and validate this comprehensive collection of data, more than a dozen textbooks, several hundred journal articles, nearly 30

Table 2.1: Comparison of the size and content of different yeast-specific or yeast-containing metabolism/metabolomics databases

Database Content	YMDB	YeastNet 4[1]	MetaCyc	KEGG
Number of metabolites	2007	792	688	720
Number of data fields	81	14	19	12
Number of NMR spectra	1540	0	0	0
Number of MS spectra	1346	0	0	0
Number of external database hyperlinks	15	4	3	5
Concentrations	Yes	No	No	No
Compound descriptions	Yes	No	No	No
Cell locations	Yes	Yes	No	No
Pathways	Yes	No	Yes	Yes
Sequence search	Yes	No	Yes	Yes
Structure search	Yes	No	No	Yes
Molecular weight search	Yes	No	Yes	No
NMR spectral search	Yes	No	No	No
MS spectral search	Yes	No	No	No
Chemical taxonomy	Yes	No	Yes	No

different electronic databases and at least 20 in-house or web-based programs were individually searched, accessed, compared, written or run over the course of the past 18 months. The team of YMDB contributors and annotators included analytical chemists, NMR spectroscopists, mass spectroscopists and bioinformaticians with dual training in computing science and molecular biology/chemistry.

The YMDB currently contains more than 2000 yeast metabolite entries that are linked to nearly 27 000 different synonyms. These metabolites are further connected to some 66 non-redundant pathways and 916 reactions involving 857 distinct enzymes and 138 transporters. More than 750 compounds are also linked to experimentally acquired 'reference' 1H and 13C NMR and MS/MS spectra. Concentration data (intracellular and extracellular) is also provided for a total of 627 compounds. The complete collection of data in the YMDB occupies a total of 1.1 GB. Relative to other yeast metabolite/pathway databases, YMDB is substantially larger and significantly more comprehensive. A detailed comparison of YMDB to other widely known yeast resources is provided in Table 2.1.

The YMDB is modeled closely after the HMDB. As a result, it has many of

[1] YeastNet 4.0 reports a total of 1494 metabolites but only 792 are unique.

the features found in the HMDB including efficient, user-friendly tools for viewing, sorting and extracting metabolites, proteins, pathways or chemical taxonomy information (Figure 2.1). These are available through the YMDB navigation bar (located at the top of every YMDB web page) that lists seven pull-down menu tabs ('Home', 'Browse', 'Search', 'About', 'Help', 'Download' and 'Contact Us'). To further aid in navigation and searching, nearly every viewable page in the YMDB, including the 'Home' page, supports simple text queries through a text search box located near the top of each YMDB web page. This text search tool, which can be specified to search through either protein or metabolite data fields, supports text matching, accommodates mis-spellings and highlights the text where the word is found. A more advanced text search that supports Boolean constructs and permits more precise data field specifications is also available.

In addition to these extensive text search capabilities, the YMDB also offers general database browsing via the 'Browse' buttons located in the YMDB menu bar. Five different Browsing options are available including Metabolite Browse (for viewing and sorting metabolites), Protein Browse (for viewing and sorting proteins), Reaction Browse (for viewing chemical reactions), Pathway Browse (for viewing yeast-specific KEGG pathways) and Class Browse (for viewing groups of compounds by their chemical taxonomy or class). Each of the Browsing views is presented as a set of navigable/sortable synoptic summary tables. These tables are, in turn, linked to more detailed 'MetaboCards' and 'ProteinCards' similar to those found in DrugBank and HMDB. Clicking on a MetaboCard or ProteinCard button opens a web page describing the compound or protein of interest in much greater detail. Every MetaboCard entry contains > 50 data fields devoted to chemical or physico-chemical data and synoptic biological data (names, sequences, accession codes). Each ProteinCard entry contains > 30 data fields devoted to biochemical, nomenclature, gene ontology and sequence data for metabolically important yeast enzymes and transporters. In addition to providing comprehensive numeric, sequence and textual data, each MetaboCard and ProteinCard also contains hyperlinks to many other databases (KEGG, BioCyc, PubChem, ChEBI, PubMed, PDB, UniProt, GenBank), abstracts, references, digital images and applets for viewing

molecular structures.

Adjacent to the 'Browse' menu, the 'Search' menu offers nine different querying tools including Chem Query, Text Query, Sequence Search, Data Extractor, MS Search, MS/MS Search, GC/MS search, NMR Search and 2D NMR Search. Chem Query is YMDB's chemical structure search utility. It can be used to sketch (through ChemAxon's freely available chemical sketching applet) or paste a Simiplified Molecular Input Line Entry Specification (SMILES) string [81] of a query compound into the Chem Query window. Submitting the query launches a structure similarity search that looks for common substructures from the query compound that matches the YMDB's database of known yeast compounds. Users can also select the type of search (exact or Tanimoto score) to be performed. High scoring hits are presented in a tabular format with hyperlinks to the corresponding MetaboCards. The Chem Query tool allows users to quickly determine whether their compound of interest is a known yeast metabolite or chemically related to a known yeast metabolite. In addition to these structure-similarity searches, the Chem Query utility also supports compound searches on the basis of molecular weight ranges.

YMDB's sequence searching utility (Sequence Search), which supports both single and multiple sequence queries allows users to search through YMDB's collection of 1104 known enzymes, transporters and other target proteins. With Sequence Search, gene or protein sequences may be searched against YMDB's sequence database by pasting the FASTA formatted sequence (or sequences) into the Sequence Search query box and pressing the 'submit' button. A significant hit reveals, through the associated MetaboCard hyperlink, the name(s) or chemical structure(s) of metabolites that may act on that query protein. With Sequence Search metabolite–protein interactions from newly sequenced yeast species or strains may be readily mapped via the *S. cerevisiae* data in the YMDB.

YMDB's data extraction utility (Data Extractor) employs a simple relational database system that allows users to select one or more data fields and to search for ranges, occurrences or partial occurrences of words, strings or numbers. The data extractor uses clickable web forms so that users may intuitively construct SQL-like

queries. Using a few mouse clicks, it is relatively simple to construct complex queries ('find all metabolites that are substrates of alcohol dehydrogenase and have boiling points above 80°C') or to build a series of highly customized tables. The output from these queries can be provided in HTML format with hyperlinks to all associated MetaboCards or as an easily downloaded comma separate value file.

YMDB's NMR and MS search utilities allow users to upload peak lists and to search for matching compounds from the database's collection of MS and NMR spectra. The YMDB currently contains 1540 experimentally obtained 1H and 13C NMR spectra (with spectral collection conditions) for 466 different compounds (most collected in water at pH 7.0, 10 mM for 1H, 50 mM for 13C) measured in our lab or obtained from the BioMagResBank (BMRB) [74]. Most of the NMR spectra are fully assigned. It also contains 951 MS/MS (Triple-Quad) spectra for 317 pure compounds analyzed by our laboratory. An additional 400 MS or MS/MS spectra were obtained from MassBank [35]. The YMDB spectral search utilities allow both pure compounds and mixtures of compounds to be identified from their MS or NMR spectra via peak matching algorithms that were developed in-house [91, 16].

Adjacent to the 'Search' menu, the 'About' pull-down menu contains information on the YMDB database, recent news or updates, links to other databases, data sources and database statistics. The 'Help' pull-down menu provides general documentation on database definitions, data field types and data field sources. It also contains information on experimental methods (for metabolite concentration measurements performed by our lab for the YMDB), details on how to cite YMDB, as well as a tutorial on how to use YMDB's advanced text search utilities. Finally the 'Download' menu contains downloadable data for all YMDB chemical structures (as Structure Data Format (SDF) files), all enzyme/protein sequences (in FASTA format), as well as complete flat file data sets of the current YMDB release in JSON format.

2.3 Database Implementation

YMDB employs a Ruby on Rails (version: 3.09)-based front-end attached to a sophisticated MySQL relational database (version: 5.0.77) at its back-end. All data are entered directly through a custom-built web interface with each YMDB MetaboCard having an edit page, which allows database curators to manually make changes to YMDB entries. The public user interface and the internal database both read from the same database.

All structures in the YMDB are stored in a centralized structure hub. This hub is a RESTful web resource that automatically stores and updates chemical properties such as molecular weight, solubility and logP. Additionally, the hub renders the structure images and thumbnails visible on the public YMDB site. The centralized nature of this structure hub helps to maintain consistency for all structures stored in YMDB. Whenever a structure is changed or updated, all properties are automatically recalculated and made available on the public site at `http://www.ymdb.ca`.

2.4 Quality Assurance, Completeness And Curation

The same quality assurance, quality control and data compilation procedures implemented during the development of HMDB, T3DB and DrugBank were used in the development of YMDB. In particular, the compounds in YMDB were identified using a combination of methods, including manual literature surveys, text mining of on-line journals or abstracts and data mining of other electronic databases. Literature sources included specialty journals on metabolomics, food composition and analysis, systems biology, analytical chemistry and textbooks on wine and beer chemistry. All primary metabolites had to have at least two databases confirm their existence and inclusion (with evidence that the necessary enzymes or pathways are present), whereas all secondary metabolites (such as those found in wine or beer) were required to have a traceable literature/experimental reference. For many secondary yeast metabolites the relevant starting compounds, reactions, pathways and catalyzing enzymes are not yet known. Hopefully, with time and improved

technology, this information will become available. With many yeast secondary metabolites it is sometimes difficult to know if the compound was present in the media (wort or grape must) prior to fermentation or whether it arose as a consequence of fermentation. For those compounds where there was some ambiguity regarding their source (plant versus yeast), we attempted to cross-check our findings through multiple literature sources in order to exclude possible grape, hops or barley metabolites.

For those yeast metabolites found to match to previously existing entries from either the HMDB or CCDB, only the chemical data fields were imported into the YMDB (except the compound description which was manually edited to include or remove organism-specific references). The biological data for these HMDB/CCDB imported compounds was generated de novo since yeast biology is very different than *E. coli* or human biology. In order to ensure both completeness and correctness, each metabolite record entered into the YMDB was reviewed and validated by a member of the curation team after being annotated by another member. Other members of the curation group routinely performed additional spot checks on each entry. Several software packages including text-mining tools, chemical parameter calculators and protein annotation tools were developed, modified and used to aid in data entry and data validation. One particular program, BioSpider [52], was used extensively to acquire routine, machine retrievable or easily calculated/verifiable chemical data on metabolites. To facilitate and monitor the data entry process, all of YMDB's data is entered into a centralized, password-controlled database, allowing all changes and edits to the YMDB to be monitored, time-stamped and automatically transferred.

2.5 Conclusions

To summarize, the YMDB is a richly annotated, web-accessible 'metabolomics' database that brings together quantitative chemical, physical and biological data about nearly 2000 *S. cerevisiae* metabolites. Relative to other yeast metabolism / pathway databases, YMDB has between 23× more metabolites and 510× more

data. The YMDB also uniquely contains detailed information on hundreds of secondary metabolites that are critically important to the food, beverage, chemical and biofuel industry. Among the other distinguishing features of YMDB are: (i) the breadth and depth of its annotations (> 80 data fields); (ii) the large number of hyperlinks and references to other resources; (iii) the availability of detailed compound descriptions; (iv) the inclusion of thousands of reference NMR and MS spectral data; (v) the inclusion of intra- and extracellular metabolite concentration data; (vi) the quantity of biological and biochemical information included in each compound entry and (vii) the support for queries by text, chemical structure, spectra, molecular weight and gene/protein sequence. Owing to these unique characteristics, we believe the YMDB fills an important niche in yeast biology as it addresses not only the specialized analytical needs of metabolomics researchers, but also the interests of molecular biologists, systems biologists, the industrial fermentation industry, as well as the beer, wine and spirit industries.

While the YMDB certainly fills an important niche for yeast metabolomics, it is also a work in progress. As with many areas in metabolomics, new compounds are constantly being discovered, new concentrations are being reported, new pathways/reactions are being elucidated and new metabolite functions are being determined. So long as our resources permit, we intend to continue to update and enhance the YMDB as this new information is published or acquired.

Figure 2.1: A screenshot montage of the YMDB showing several of the YMDB's search and data display tools describing the metabolite L-Glutamine. Not all fields are shown.

Chapter 3

SMPDB 2.0: The Small Molecule Pathway Database

3.1 Introduction

Biological pathways are the wiring diagrams of life. They provide a rich and surprisingly compact view of how genes, proteins and metabolites work together in cells, tissues or organs. In fact, most of today's life scientists learned their biochemistry and molecular biology by studying richly annotated, carefully hand-drawn pathway diagrams found in books or on wall charts. With the advent of the internet, many biological pathway diagrams started migrating from the printed page to the web. As a result, there is now a plethora of very popular, very high-quality web-accessible pathway databases such as KEGG [45], the 'Cyc' databases [49, 6], the Reactome database [11], WikiPathways [63] and PharmGKB [73]. The advantages of using the web to illustrate and disseminate biological pathway information are manifold. These include greater accessibility, improved interactivity and enhanced functionality. However, the limitations associated with creating hyperlinked web illustrations often means that visual interactivity has to come at the expense of artistic quality or biochemical detail. This can be particularly problematic in the fields of clinical metabolomics and clinical chemistry where details about chemical structures as well as target organs, tissue locations, organelle function and physiological activity are particularly important.

In an effort to make web-based biological pathway data more visually appealing and physiologically relevant, particularly for biomedical applications, we developed

the Small Molecule Pathway Database or SMPDB [20]. The first version of SMPDB, which was released in 2010, contained > 350, colorful, artistically rendered and biochemically complete pathway illustrations describing many key aspects of human metabolism. These included hyperlinked diagrams of human metabolic pathways, metabolic disease pathways, metabolite signaling pathways and drug-action pathways. Many of these pathway diagrams included information on the tissue locations, relevant organs, organelles, subcellular compartments, protein cofactors, protein locations, metabolite locations, chemical structures and protein quaternary structures. As with many high-quality web-based pathway databases, SMPDB also provided extensive hyperlinking, browsing, searching and annotation functions along with detailed pathway descriptions and references.

However, the first version of SMPDB was not without some shortcomings. In particular, it was difficult to update and maintain, it was limited in size and scope, its pathways lacked a certain degree of standardization, its physiological information on organelles and organs was incomplete and its visual interactivity (zooming and image navigation) was awkward and somewhat restricted. Furthermore, SMPDB pathway diagrams were only downloadable as static images and not available in standard formats such as BioPAX or SBML [71]. While SMPDB did link to well-known databases such as DrugBank [51], HMDB [86] or UniProt [75], these databases did not link to it. In other words, SMPDB was not particularly visible to the metabolomics, drug research or systems biology communities.

With the release of SMPDB 2.0, we believe we have addressed all of these shortcomings. In particular, we have developed tools to simplify SMPDB's maintenance and enhance its illustration standards. We have also corrected and added much more physiological (tissue, organ, organelle and transporter) information, substantially improved its visual interactivity and quality, created downloadable BioPAX pathway files for all of SMPDB's pathways and increased SMPDB's visibility by integrating SMPDB into DrugBank, HMDB and MetaboAnalyst. Finally, we have substantially expanded the size and scope of SMPDB to include > 600 pathways (a 70% increase) and have added a large number of drug metabolism and physiological action pathways. With these enhancements, we believe that SMPDB 2.0 will

appeal to a much broader community of researchers and will offer users a much more enriched, interactive and informative experience. A detailed description of SMPDB 2.0 follows.

3.2 Increased Size And Scope

When it was initially released, SMPDB (version 1.0) contained 350 hand-drawn pathways. These small molecule pathways were limited to three general categories: basic metabolism, metabolic diseases and selected drug-action pathways. Today, SMPDB 2.0 contains > 600 pathways (a 70% increase) and it now includes six kinds of small molecule pathway categories including: basic metabolism, metabolic diseases, drug-action pathways, drug metabolism pathways, metabolite signaling pathways and physiological action pathways. The inclusion of drug metabolism pathways in SMPDB 2.0 was driven by user requests and the rapidly growing interest in drug metabolism in many fields of drug research and clinical metabolomics. The addition of metabolite signaling and physiological action pathways was motivated by the need to handle a larger number of drug-action pathways. It was also brought on by the growing awareness by members of the metabolomics and nutritional science communities that many metabolic pathways act at the level of organs and tissues and not only within cells. In addition to the significant numerical increase in SMPDB's size and scope, many of the pathway diagrams in SMPDB 2.0 have been enhanced with additional cellular or physiological information. Now nearly every SMPDB pathway includes information about where the reactions occur (cellular compartment(s), organs or tissues), the site/organ of action for drugs (for drug-action pathways) or toxic metabolites (for metabolic disorders) as well as information on key membrane-bound transporters (which move drugs and metabolites in and out of almost every cell). This kind of physiological or extracellular information is rarely captured or displayed in other online pathway databases. However, because SMPDB is a human-only pathway resource, this sort of physiological, cellular, biochemical or biomedical information is particularly important. Consequently, considerable effort over the past 3 years has been put into capturing

and displaying these data.

As with its predecessor, SMPDB 2.0 is still strictly focused on curating human-only, small molecule-only pathways. Therefore, SMPDB only displays those pathways where small molecules (metabolites or drugs (< 1500 daltons) play key roles and/or where small molecules represent a significant proportion of functional entities in a given pathway. All of the pathways in SMPDB 2.0 still retain the standard (and, in some cases, unique) database functionalities of the original database, including hyperlinked structures (to HMDB or DrugBank), hyperlinked protein images (to UniProt), text searching utilities (TextQuery), chemical searching utilities (ChemQuery), sequence searching utilities (Sequence Search), pathway browsing capabilities (SMP-Browse with filtering options), detailed pathway descriptions, extensive references, metabolite highlighting capabilities (SMP-Analyzer and SMP-Highlight), pathway legends and metabolite mapping for metabolomic applications (SMP-MAP). Additional details about these functions and how they can be used are provided in the original SMPDB paper. An updated and detailed comparison of SMPDB 2.0 to its predecessor (SMPDB 1.0) and to other pathway databases is provided in Table 3.1.

3.3 Enhancements In Visualization And Interactivity

One of the distinguishing features of SMPDB has been the strong focus on providing users with high-quality, artistically pleasing pathway diagrams that were not only correct and informative but also colorful, interactive and richly detailed. This continues to be a major focus for SMPDB 2.0. To this end, a significant effort over the past year has been put into enhancing the visual displays and interactivity in SMPDB. The previous version of SMPDB only offered a three-step zooming capability that was further hindered by the requirement that users navigate through the zoomed-in images using multiple scroll bars. This restricted zooming capability proved to be both awkward and inconvenient for many users. Likewise, the placement of the pathway descriptions and references (above and below the pathway images) limited how much of a pathway could be viewed on a web page or

Table 3.1: Comparison of SMPDB 2.0 to SMPDB 1.0 to KEGG, HumanCyc, Reactome, BioCarta and WikiPathways/GenMAPP

Feature	SMPDB 2.0	SMPDB 1.0	KEGG	Reactome	HumanCyc	BioCarta	WikiPathways
Number of metabolic pathways	92	70	78 (for humans)	64 (for humans)	333 (short pathways)	69	50 (for humans)
Number of disease pathways	221	113	52	11	0	24	16
Number of drug action pathways	232	168	0	3	3	10	5
Number of drug metabolism pathways	53	0	8	1	0	0	2
Number of physio. action pathways	5	0	31	36	0	5	22
Number of small mol. signaling pathways	15	0	30	27	0	30	15
Provides multiple organism pathways	No	No	Yes	Yes	No	Yes	Yes
Chemical structures shown in diagrams	Yes	Yes	No	No	Yes (when zoomed)	Some	No
Protein 4° structures shown in diagrams	Yes	Yes	No	No	No	Some	No
Cell structures shown in pathway diagrams	Yes	Yes	Some	Yes	No	Yes	No
Organs shown in pathway diagrams	Most	Some	No	No	No	No	No
Descriptions of pathways provided	Detailed	Detailed	Limited	Detailed	Detailed	Detailed	Limited
Pathway images are hyperlinked	Yes	Yes	Yes	Yes	Yes	Yes	No
Pathway images are easily zoomable	Yes	No	No	Yes	Yes	No	Limited
Information provided on pathway entities	Detailed	Detailed	Modest	Modest	Moderate	Limited	Limited
Supports advanced text search	Yes	Yes	No	Yes	Yes	No	No
Supports sequence searching	Yes	Yes	No	No	Yes	No	No
Supports graphical chem. structure search	Yes	Yes	No	No	No	No	No
Supports chemical expression mapping	Yes	Yes	Yes	Yes	Yes	No	No
Supports gene/protein expression mapping	Yes	Yes	Yes	Yes	Yes	No	No
Downloadable	Yes	Yes	Limited	Yes	Yes	No	Yes
BioPax, CellML or SBML compatible	Yes	No	Partial	Yes	Yes	No	No

computer screen. In addition, the size of the chemical structures (too large), the text labels (too small) and the pathway arrows (too thin) was often problematic for a number of pathways, depending on what level of zooming was used. The uniformly dark blue background (to indicate an aqueous environment) was also challenging for those wishing to print SMPDB pathways on paper, for preparing slides and for visualizing certain features or distinguishing among subcellular locations.

For SMPDB 2.0, a completely new visualization engine was built using scalable vector graphics (SVG) and a web interface technology inspired by Google Maps (`http://en.wikipedia.org/wiki/Google_Maps`). This enhancement now allows rapid and continuous zooming using a mouse scroll wheel or through simply clicking on-screen zoom icons. It also allows facile navigation around zoomed-in pathway diagrams in SMPDB 2.0 through a simple click-and-drag operation or through clicking on-screen up/down or left/right arrows located near the on-screen zoom functions. A full-screen view of each SMPDB map is also available. This view can be toggled off and on by clicking the full-screen icon located on SMPDB's new information and control panel (located on the right side of each pathway map). The information and control panel allows users to easily click on labeled buttons to view text (pathway descriptions and references), highlight (SMP-Highlight), annotate concentration data (SMP-Analyze), download files and adjust image settings. By moving the old text boxes containing pathway descriptions and pathway references to SMPDB 2.0's information and control panel, more on-screen real estate is now available for viewing pathway diagrams. Furthermore, by resizing, recoloring and standardizing the size of the chemical structures, arrows and text in all of SMPDB's pathway images, most features under most zooming conditions should be much more visible and easily discerned. Using the image settings located in the control panel, users can also adjust the background colors (from blue to white) and the cellular membrane display (simple versus detailed) to facilitate printing, image capture or slide preparation. The recoloring process has also been extended to other parts of SMPDB as well. In particular, all reactions that take place in certain organelles or subcellular locations have now been recolored so that the organelle's background color serves as the background color for the (zoomed-in) reaction. This

should make compartment-specific reactions or pathways far more distinctive and discernable. Another enhancement to SMPDB is now available through the home page where a scrollable 'carousel' of SMPDB pathway images is available. This carousel, which can be scrolled through by simply 'swiping' the mouse over the images, will allow users to view and select newly loaded or recently updated SMPDB pathway maps. A montage of many of these visualization features and enhancements is shown in Figure 3.1.

3.4 Improved Standardization And Reduced Maintenance

SMPDB was originally built by hand-drawing each pathway using PowerPoint incorporating a collection of visual icons and a well-defined standard operating protocol (SOP). Each PowerPoint image was then converted to three differently sized PNG images (small, medium and large) and manually image-mapped to appropriate database links. Prior to the image mapping process, pathway constituents (proteins and chemicals) were manually identified, their database IDs manually determined and each data point entered in a separate file for each SMPDB pathway. These text files were used to permit text and structure searching as well as compound highlighting (via SMP-MAP and SMP-Highlight). This manually intensive process proved to be arduous, time-consuming and error prone. The many different skills and multiple steps required to generate and post a SMPDB pathway meant that a team of more than 10 differently trained individuals were needed to construct, update and maintain the database. Even with SOPs and a standardized icon library, some individual variability occurred with regard to pathway layout, pathway details, icon placement, directional arrow size and other image features. Additionally, the manual placement of image-mapped hyperlinks meant that some links were placed imprecisely, leading to off-centre effects that became further exacerbated through zooming. If SMPDB pathways needed to be updated or corrected, multiple images still had to be generated and multiple image maps still had to be prepared. If a large number of highly similar pathways were involved (such as drug mechanism path-

ways) a simple, single molecule update could often take many hours of repetitive, error-prone editing. Because all of SMPDB's pathways were essentially image-mapped pictures it was not possible to easily render them into BioPAX or SBML formats. As a result, efforts to convert SMPDB pathways into BioPAX or SBML equivalents were stymied.

To simplify the updating process and improve the level of standardization and figure rendering consistency for SMPDB 2.0, we developed an online pathway illustration and editing tool, called PathWhiz (manuscript in preparation). PathWhiz, which is a combination of an online digital painting program and a webpage building program, allowed SMPDB curators to digitally render, update and annotate pathways in a fraction of the time that the old 'analog' approach required. PathWhiz also allowed our curation team to produce far more consistently rendered pathways and consistently sized structures that looked as if they all came from a single artist. PathWhiz uses a pull-down library of standard image icons for membranes, membrane compartments, organs, tissues, organelles, membranes and proteins to help accelerate the pathway illustration process. It also has a standard set of arrows and arrow-drawing utilities that allow all SMPDB reaction components to be linked or connected in a uniform, visually pleasing fashion. Because PathWhiz also captures icon placement and annotation information digitally, it allowed all hyperlinks to be precisely mapped, all pathway images to be rendered as SVG and all pathway component information (names, structures, database IDs, etc.) to be readily captured. The updating process is also made much easier as all changes in one pathway (images, hyperlinks, names, database IDs) can be digitally transferred to all other similar pathways at the click of a button. The digital nature of the drawing and icon placement process through PathWhiz meant that it also became very easy to convert all SMPDB 2.0 pathways into BioPAX format. Conversion of the SMPDB pathways to SBML is underway now and should be completed before 2014. Thanks to PathWhiz, all SMPDB pathways have been significantly updated, corrected and improved. Likewise many new pathways were able to be added in a fraction of the time (compared to the old approach), allowing a significant increase in both the quantity and quality of pathways in SMPDB 2.0.

3.5 Enhanced Data Downloads

Another important, and increasingly unique, feature of SMPDB is the open availability of its data. In particular, SMPDB supports both data downloads and image downloads for all of its pathways. The data files include protein and gene sequences for all the annotated proteins and genes in SMPDB. SMPDB downloads also include all metabolite and drug structures (in SDF format) as well as CSV files containing different combinations of compound names or protein names linked to pathways and database identifiers. The image files (in SMPDB 1.0) included both the original PowerPoint files and the differently sized (small, medium, large) PNG files corresponding to each pathway. With the release of SMPDB 2.0 the same kind of data download information is still available, although the number of sequences, structures and database links is now substantially greater. In addition, SMPDB's data downloads now include all of its pathway information in BioPAX format. BioPAX is an RDF/OWL-based standard exchange language designed to compactly represent biological pathways at the molecular and cellular level [79]. The availability of SMPDB's data in BioPAX format and the impending availability of the same pathway data in SBML (expected in late 2013) should greatly expand SMPDB's appeal to systems biologists as well as its potential utility in a variety of pathway, biochemical and metabolomic analysis packages. It is also expected that the availability of SMPDB BioPAX (and SBML) files will encourage greater interest in community annotation, allowing SMPDB to evolve from a centrally curated database (which is inefficient) to one that may be curated and enhanced by a community of experts (which is more efficient). SMPDB 2.0's image file downloads have also been modified and enhanced to include not only high-resolution PNG files but also a SVG images of every pathway with all of the BioPAX data imbedded (as metadata) into every file. With these enhancements to SMPDB 2.0's downloadable resources, we believe it now offers the most comprehensive collection of downloadable data of any online pathway database.

3.6 Improved Quality And Quality Assurance

As noted earlier, every pathway in SMPDB 2.0 has been redrawn using the new PathWhiz rendering tool. This effort allowed us to revisit all of SMPDB's pathways and to update them with new information and to correct, improve or reformat preexisting pathways so that they were more consistent and informative. Many pathways were improved through the addition of more organelle and reaction compartment information. Others were enhanced through the addition of target organ information and protein transporter information. Likewise images that were too small, arrows that were too thin and pathways that were too convoluted were corrected and/or simplified. In redrawing old pathways and rendering new pathways in SMPDB 2.0, the same level of quality assurance, the same referential resources and the same kinds of data checking/validation routines were used as described in the original SMPDB paper [20]. In particular, all pathways in SMPDB 2.0 were drawn using a standard operating procedure (SOP) with a checklist of features that each of the pathway artists/annotators was required to follow. The SOP and checklist are also provided in the 'About' menu. To ensure that the pathways have been knowledgably illustrated, all of the SMPDB pathway curators were required to have degrees or advanced university coursework in biology, chemistry, biochemistry or bioinformatics. During the redrawing of existing SMPDB pathways and the generation of new pathways for SMPDB 2.0, all SMPDB curators were also required to consult a variety of sources including biochemistry textbooks, OMIM [29], Wikipedia, KEGG [45], MetaCyc [6], Reactome [11], UniProt [75], the HMDB [86], DrugBank [51] and PharmGKB [73]. This allowed the curation team to identify and consolidate pathway nomenclature, key pathway components, critical reactions, cellular compartments as well as target organs, compartments or organelles. Pathway layouts were also frequently assessed, compared and discussed by SMPDB team members prior to manually generating any pathway diagrams. Because of the unique rendering requirements and the strict SOPs for SMPDB, no pathway in SMPDB was 'copied' from any other pathway diagram in any other database. Furthermore, SMPDB's metabolic disease pathways, drug-action and drug metabolism pathways had

to be generated independently of any pathway database. Instead, relevant information was gathered from various medical and pharmacology textbooks, specialized encyclopedias and fragmentary data contained on several online databases such as OMIM [29], PharmGKB [73] and DrugBank [51]. As an additional layer of quality assurance, all of the pathway diagrams in the SMPDB 2.0 have been inspected and corrected by two or more curators having advanced degrees in biochemistry or physiology.

3.7 Increased Connectivity And Interoperability

When SMPDB was first released, it was primarily a standalone database that simply accessed other databases through its hyperlinks to HMDB, DrugBank and UniProt. Over the past 2 years, efforts have been directed at integrating SMPDB into a number of other popular online databases, so that effectively other databases now access SMPDB. In particular, SMPDB is now linked into the pathway data fields of the latest version of HMDB [86] and the latest release of DrugBank [51]. It has also been linked into several online metabolomic data analysis resources including MetaboAnalyst [92] and MSEA [93]. Discussions are ongoing to have SMPDB 2.0 linked to a number of other chemical entity databases as well as other well-known pathway databases in the near future. In addition to trying to increase SMPDB's database connectivity, we are also trying to make it far more interoperable. The availability of SMPDB's data in BioPAX format and the expected availability of the same pathway data in SBML (in late 2013) should greatly improve SMPDB's interoperability. It should also encourage users to freely exchange SMPDB data and to incorporate these data into a variety of pathway or metabolomic analysis software packages. It is also expected that the availability of SMPDB BioPAX (and soon SBML) files will encourage greater interest in community annotation, allowing SMPDB to evolve from a centrally curated database to a community curated resource.

3.8 Future Plans And Conclusions

SMPDB continues to expand with new pathways being added at an almost daily rate. Over the coming 2–3 years, we expect that the number of pathways in SMPDB will likely double. It is likely that the greatest growth in SMPDB's pathways will be in the drug and drug metabolism pathways, as there are literally thousands of reactions, reactants and pathways that are readily available in the literature. With the impending release of PathWhiz (SMPDB's pathway drawing tool and editor), we are hopeful that many of SMPDB's new pathways will actually be added by interested community members. It is also expected that PathWhiz will enable SMPDB-like pathways and pathway databases to be easily generated for other model organisms such as *Saccharomyces cerevisae*, *Escherichia coli*, *Arabidopsis* and *Drosophila*. With the latest additions and enhancements in SMPDB 2.0, we believe that SMPDB has reached a critical threshold giving it sufficient breadth, depth and interconnectivity so that it will appeal to a much larger community of users. Overall, SMPDB 2.0's new, colorful, informative, artfully designed pathway diagrams, combined with its wide range of visualization, annotation and querying tools should provide users a much more enriched, interactive and informative pathway viewing experience.

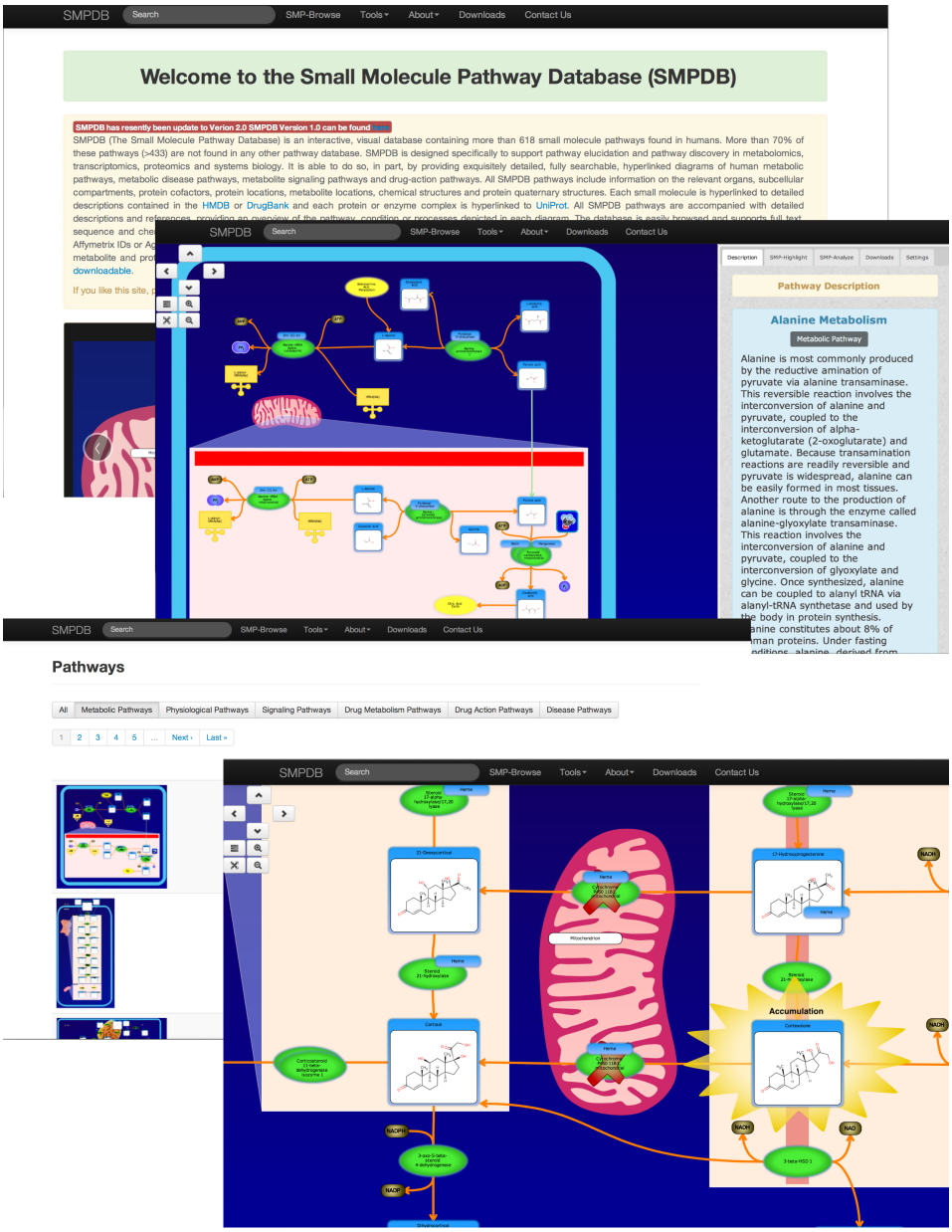

Figure 3.1: A screenshot montage of SMPDB 2.0's various viewing and searching features.

Chapter 4

PathWhiz: a Web Server for Pathway Generation and Visualization

4.1 Introduction

Pathway diagrams are the roadmaps of biology. They have long been used as visual tools to map out complicated biological processes over space and time. In their simplest form pathway diagrams can be used to illustrate the connections between genes, proteins and/or metabolites — especially at the cellular or subcellular level. More complicated pathway diagrams attempt to extend these connections to cells, tissues, organs or entire organisms, thereby adding even greater biological context. Regardless of whether the actual pathways are simple or complex, the goal of a pathway illustration is to render complex biological processes and connections in a way that a viewer can easily understand.

There is no single correct way to illustrate a biological pathway. Indeed, the level of detail and the amount of information conveyed for identical biological pathways or even within components of the same pathway can vary tremendously. These differences often depend on the intended purpose or the state of knowledge of a given pathway. Certainly when pathway diagrams are created for textbooks or wall charts, the pathway illustrator is expected to generate complex, richly detailed, colourful works of art [58]. In these artistic renderings the structures of metabolites, proteins (if known), DNA, cell membranes, cell structures, cell organelles and even

relevant organs are often shown along with labels, directional arrows and detailed notes or commentaries. However, when pathway diagrams are generated for internet applications this kind of artistry and much of the biological context is typically sacrificed in favour of generating machine readable wire diagrams or schematics. The main redeeming feature of these simplified internet pathway diagrams is that they are often richly hyperlinked and are interactively zoomable. Many popular web-based pathway databases, such as KEGG [44], MetaCyc [46], Wikipathways [48], and Reactome [57] contain these kinds of simplified pathway diagrams.

More recently, a number of these pathway databases have made their software available to allow others to illustrate or generate web-compatible pathway diagrams. In particular, BioCyc's Pathway Tools [47], and Wikipathway's PathVisio software [77] let users generate and share biological pathways using downloadable software. Wikipathways also offers an online wiki-style tool for the generation of biological pathways [48]. They also allow users to generate pathways in a machine readable BioPAX format. In addition to these database-associated pathway rendering tools, there are also a number of commercial tools and stand-alone freeware packages for generating biological pathways including Cytoscape [67], CellDesigner [21], PathCase [17] and VisANT [36]. However, just as with PathVisio and BioCyc's Pathway Tools, these software packages are limited to generating very simplified, highly schematic pathway diagrams that tend to be more pleasing to computers than the human eye.

Ideally what is needed is freely available, user-friendly software system that provides the capacity to generate colourful, visually pleasing and biologically complete pathway diagrams found in textbooks while at the same time supporting the generation of machine readable, interactive, richly hyperlinked pathway diagrams that are compatible with internet databases. The fact that no such software system exists led us to develop a novel kind of pathway drawing tool called PathWhiz.

PathWhiz is essentially a web server designed for the facile creation of colourful, visually pleasing and biologically accurate pathway diagrams that are machine readable, interactive and fully web compatible. PathWhiz differs from other pathway drawing tools in that it is a web server rather than a stand-alone program. This

makes PathWhiz accessible from almost any place and compatible with essentially any operating system. PathWhiz also differs from most other pathway drawing tools with regard to the level of biological detail, physiological context and biological complexity that it can support. In particular, PathWhiz, through a specially designed drawing palette, allows the facile rendering of metabolites (via automated generation of their structures), proteins (including quaternary structures), covalent modifications, cofactors, membranes, subcellular structures, cells, tissues and organs. Furthermore, PathWhiz has been designed so that the construction of pathways and the hyperlinking of pathway features can be done quickly and intuitively. Additional details and specific examples regarding the design and implementation of PathWhiz are given below.

4.2 Implementation

PathWhiz consists of three major components: a Pathway Editor, a Pathway Viewer and a PathWay data repository for capturing metadata about each pathway and pathway object or process. Pathways are generated and edited using the Pathway Editor while the Pathway Viewer is for visualizing, printing or downloading the finished pathways. PathWhiz was built on a Ruby on Rails (version 3.2.13) web framework incorporating a MySQL relational database (version: 5.0.77) to manage all of the pathway data, including entity relationships, URLs, descriptions, object images and chemical structures. The front-end web client is controlled by Ruby on Rails combined with 'Backbone.js' (version 1.0.0) as the front-end web framework. PathWhiz collects and manages all the data necessary to produce scalar vector graphic (SVG) images annotated with BioPAX [14] representations of the pathway data. PathWhiz has been tested and found to be compatible with most modern web browsers including: Google Chrome (v. 31 and above), Internet Explorer (v. 9 and above), Safari (v. 7 and above), Opera (v. 15 and above) and Firefox (v. 23 and above).

4.3 Pathway Generation and the Pathway Editor

In PathWhiz all biological pathways are modelled using three standard elements: biological objects, biological processes and meta-data about these objects or processes. Objects are typically proteins/enzymes, metabolites, cofactors, membranes, organelles and organs, while processes are reactions, binding events and transport activities. The meta-data about objects and processes may include object names, accession numbers, structures, URL's, reaction/transport directionality and other kinds of notes or annotations. In PathWhiz, both objects and processes (as well as their associated meta data) are entered, edited and positioned using PathWhiz's Pathway Editor (see Figure 4.1 and Figure 4.2). The Pathway Editor is a specially designed, online graphical workspace where pathway objects (proteins, metabolites, co-factors, etc.) and processes (reactions, transport events, binding events) can be rendered and manipulated to create the desired pathway diagram. Individual pathway elements (both objects and processes) may be selected from the Pathway Editor's pull-down menus. Currently the menu includes: Create new reaction, create new transport, create new reaction coupled transport, create new membrane, create new zoom box, create new sub-pathway, create new image and create new label. Once an object or process is selected from the menu, the user may use the Pathway Editor to position these elements using their mouse to click and drag them to their desired position in the drawing palette space. The Pathway Editor also uses the standard 'click and drag' mechanism for selecting multiple pathway objects for group positioning or minor group adjustment. The orientation of pathway objects relative to their connecting elements (i.e. processes) is fully controlled by the user. Unconnected elements are always placed in the upper left corner of the drawing palette until connected. Each object or process that is pasted or placed into the Pathway Editor drawing palette is automatically saved to that pathway diagram. Once a pathway diagram is completed or even if it is only partially finished, it can be instantly viewed through PathWhiz's Pathway Viewer (see below). This viewer provides an intuitive zoomable, draggable, interactive visualization interface for viewing or downloading the pathway image and its associated data.

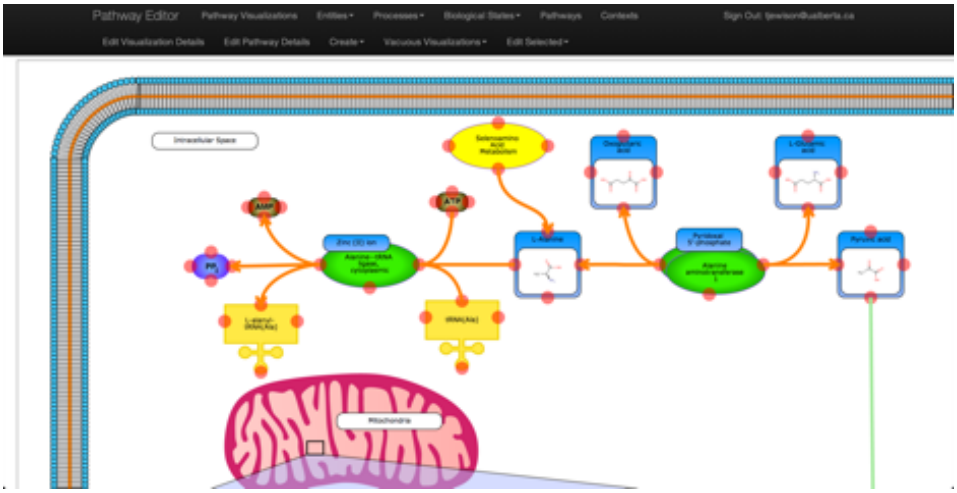

Figure 4.1: The PathWhiz Pathway Editor. Top menu bar provides quick access for all tools to create and render any pathway elements. Pathway visualizations are manipulated through the main window where elements can be placed and dragged into position.

PathWhiz uses a series of web forms to capture all of the necessary metadata needed to describe and annotate a pathway. To create a new pathway the user must first complete a web form providing the pathway name and a short description. Once that is completed the user is then taken to the Pathway Editor. Once in the Pathway Editor, all objects and processes associated with the pathway being rendered can be selected from the Pathway Editor menu (located at the top of the window see Figure 4.1). If one wishes to draw a specific enzymatic reaction, a biological state representing the spatial location of that reaction is first selected and then the appropriate enzymes and reactants (i.e. metabolites) are specified. Each of the selected objects can be customized from a standard set of templates through a web form before final rendering or placement in the Pathway Editor's drawing palette. Users may change the size (stepwise for some objects, freely scalable for others) and orientation of reaction arrows, transport arrows, proteins, drugs, cofactors, organelles and organs through click and drag operations as well as menu buttons for scaling operations. Each process (reaction, transport event, binding event) in the pathway can be extended by selecting an element or elements at the beginning or end of the particular process being edited. This operation is continued until all desired reactions or pathway elements are completely connected, fully rendered

Figure 4.2: The PathWhiz Pathway Editor Menus. The PathWhiz menus provides links to the forms for the creation of biological representations including compounds, nucleic acids proteins, polypeptides, tissues, cell types, subcellular locations, reactions and transportation. The menus also include links for creating new visual representations and editing the meta-data details of a pathway.

and properly placed in their desired positions.

Throughout the pathway rendering process users must provide information (i.e. compound names, UNIPROT accession numbers, HMDB metabolite IDs) about the pathway elements using PathWhiz's web forms. In many cases the provision of a UNIPROT accession number or an HMDB ID is sufficient for the program to fill in the remaining information (structure, name, cofactor state, hyperlink URL, etc.). In all cases, the data specifications and data entry formats used by PathWhiz are designed to meet standard BioPAX 3.0 requirements. BioPAX is a widely adopted data exchange format developed specifically for sharing biological pathway data [14]. This makes it ideal for making PathWhiz's pathway data machine readable and readily exchangeable. PathWhiz currently does not support the full breadth of data representation available in BioPAX 3.0 but it covers essentially all of what is required to represent most kinds of pathways (metabolic, signalling, protein-protein interaction) with their tissue, organ and subcellular specificity. This includes biochemical reactions, covalent modifications, non-covalent binding and the transport of molecules between different locations. The use of BioPAX to capture and store all of the pathway data in PathWhiz helps ensure that all pathway diagrams pro-

duced by the Pathway Editor are of high quality and are fully supported with verifiable and standardized data formats. Adopting BioPAX also means that PathWhiz can support the reading or uploading of any previously existing PathWhiz pathway. It also means that the PathWhiz Pathway Editor can be used to read, edit or re-render any other pathway generated by other software tools and saved using the BioPAX data exchange format.

PathWhiz supports the rendering of chemicals (metabolites), nucleic acids and monomeric proteins as well as the drawing of higher order structures (quaternary or subunit structures for proteins) along with covalent chemical modifications. PathWhiz can also represent non-covalently bound pathway objects, which may correspond to protein-DNA complexes or receptor-ligand complexes. All these pathway objects can be cross-referenced with additional information regarding their known biological states within a given organism, tissue or cell type. Simple biological processes are limited to chemical reactions (transformation and covalent modification), non-covalent binding, direct transport mechanisms and chemically-coupled transport mechanisms.

PathWhiz also supports a membrane drawing tool that provides templates for generating membrane boundaries or compartment barriers corresponding to nuclear, mitochondrial or plasma membranes. A membrane can be drawn as a rounded-rectangular enclosure or as a free form line-like drawing. The membrane can be further manipulated through control points that allow one to adjust its position and curvature. Pathways can be further augmented with stock images corresponding to other kinds of biological objects such as the liver, kidney, brain, mitochondria or nucleus amongst many others. These pre-rendered images can be resized and their positions adjusted using click and drag operations. These biological objects are typically used in conjunction with a generalized container called a 'Zoom Box' that can be used to illustrate the containment of entities within a particular biological location. Zoom boxes are edited through control points that can be used to adjust their size and position.

The PathWhiz website (http://smpdb.ca/lims) also provides a user manual and a short visual tutorial showing how a simple pathway diagram can be generated

or edited. Because PathWhiz operates as a web server, all pathways generated by any of its users are automatically stored on the web server's disk. If users choose to make their pathway diagrams public or if they choose to submit them to one of PathWhiz's associated pubic databases (i.e. SMPDB) they can do so by clicking on one of the pathway submission options. If users choose to keep their pathways private they must obtain a PathWhiz account. Users can also install PathWhiz locally and information about its operating system compatibility and directions for its download and complete installation are available on the PathWhiz website.

4.4 Pathway Viewer

PathWhiz's Pathway Viewer (see Figure 4.3) provides a convenient interface for viewing and manipulating all of PathWhiz's pathway diagrams. Using this viewer a pathway diagram can be easily zoomed in and out as well as navigated in a similar fashion to Google Maps. All of these operations are performed using standard navigation buttons and conventional mouse-activated click and drag operations. We believe this type of image navigation is very appropriate for viewing biological pathways since they are essentially road maps for biologists. The PathWhiz Pathway Viewer also provides a simple highlighting functionality that allows users to quickly find and jump to specific entities of interest. Simply entering the name of the object or entity of interest in the Search textbox and clicking the return key is all that is needed.

Most elements within a PathWhiz pathway are hyperlinked and, when clicked, will provide the user with a pop-up window giving a synoptic view of the selected element, its participating processes and its biological state information. The Pathway Viewer also allows the user to enter, annotate and visualize experimental metabolomic or proteomic data by mapping metabolite/protein concentrations to a colour gradient that is applied to the annotated elements. Through the Pathway Viewer users may also download pathway diagrams in a variety of formats for further offline use. In particular, pathway diagrams can be downloaded as SVG (support vector graphic) images, as PNG (portable network graphic) images or as

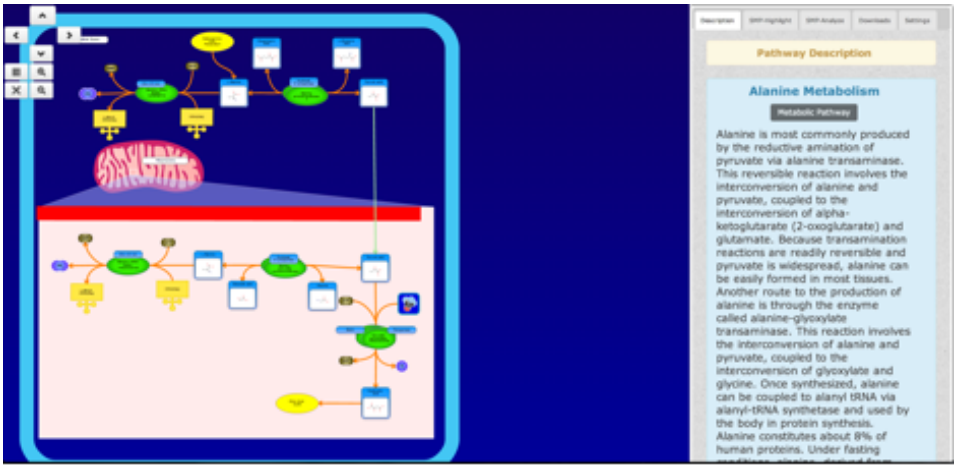

Figure 4.3: The PathWhiz Pathway Viewer. The top left interface buttons provide basic navigation, zooming, full screen mode and hide side menu actions. The central view port displays the pathway which can be navigated by click and drag as well as zoomed using the mouse. The right menu bar displays the description of the pathway with associate references as supplied by the user. The menu also provides searching and highlighting features through the SMP-highlight menu. SMP-Anaylze provides forms to input experimental concentration data to be mapped to the pathway. Downloads offers links to image and BioPax format downloadable files. Settings allows for visual customization.

BioPAX files. The SVG image actually captures all of the metadata used to create the pathway in BioPAX format which is stored with the SVG image. The PNG image is a pure image file and does not have any associated metadata. These PNG images are typically intended for printing, presentations or publications. Likewise, the BioPax download contains only the BioPax text data and has no image data. The pathway viewer allows for the changing of the background color to white to aid in printing of pathways. As well, the membrane visualizations can be toggled between a stylized lipid bilayer (better looking but slower performance in view) and simplified line representation (less attractive but offers better performance on large pathways in the viewer) of the membranes.

4.5 Discussion

Pathway diagrams drawn by skilled artists are typically the most aesthetically pleasing and most easily understood of all forms of pathway visualizations. Hand-drawn

pathways give the illustrator/scientist tremendous freedom to represent a pathway in a way that best communicates their understanding of the biology and the pathway being depicted. However, if one is interested in drawing dozens or hundreds of pathways or if one is wanting pathway illustrations to be compatible with the internet (i.e. zoomable, navigatable, hyperlinked), then relying on a single artist or giving one or more pathway illustrators complete artistic freedom can lead to significant problems with productivity, internet compatibility, quality, consistency and interpretability. These problems exist because drawing pathways is just as much an art form as it is scientific pursuit. Each artist has their own style that style can change or develop over time and even vary depending on the subject. Certainly many problems with consistency can be mitigated by the use of standard operating procedures (SOPs), but even SOPs will not eliminate the need for external reviewers and editors to ensure pathways meet quality specifications.

These problems are not unique to hand-drawn pathway databases. Issues of quality, consistency and completeness seem to plague even the best internet pathway databases even with their simplified wire-frame schematic diagrams. Nevertheless, these were precisely the problems we experienced when we tried to develop a database of small molecule pathways called SMPBD in 2009 [20]. The goal of SMPDB was to provide internet accessible versions of high-quality hand-drawn pathways. While the results were visually appealing and the database proved to be popular within the target community, the use of multiple pathway artists led to problems with pathway consistency and quality. Furthermore the reliance on static images and hand-annotated image maps prevented pathways in the database from being updated or easily edited.

It was from the many frustrating experiences associated with the construction and maintenance of SMPDB that we decided to develop PathWhiz. Essentially, PathWhiz is an attempt to combine the advantages of software-rendered pathway drawings with the aesthetic appeal of higher quality hand-drawn pathway diagrams. In effect PathWhiz incorporates well-defined SOPs into a pathway drawing program, thereby creating an environment for the generation of consistent and aesthetically appealing pathway drawings supported by rigorous data collection standards.

While far from perfect, we believe PathWhiz addresses many of the problems and shortcomings of our previous pathway efforts. It also offers a number of important features and has some significant advantages over other pathway drawing tools. First, PathWhiz supports the generation of colourful, complex, visually pleasing and biologically rich pathway diagrams in a highly standardized way. Second, it allows for the simple and rapid generation of image-mapped pathway diagrams that can be easily zoomed, clicked or navigated via the internet. Third, it provides a framework for reading, writing and rendering pathway diagrams in a machine readable, easily exchanged format (BioPAX). Fourth, as a web server, PathWhiz makes biological pathway generation far easier and pathway accessibility much greater because it is largely platform-independent and accessible from almost anywhere, anytime. Fifth, because of its web-server design, PathWhiz supports community pathway contributions. (similar to Wikipathways) This opens the door to a form of pathway crowd-sourcing to more rapidly generate new pathways for database deposition or sharing.

PathWhiz has been extensively tested by nearly a dozen different users for more than a year. Throughout this testing period it has been extensively optimized and its operations and GUI streamlined through constant user feedback. As a result of these improvements, PathWhiz was recently used to create one of the world's largest and most complete collections of pathway diagrams [41]. In particular, PathWhiz was used to generate more than 600 different human metabolomic pathways in SMPDB (version 2.0) covering common metabolic pathways, metabolic disease pathways, signalling pathways as well as drug metabolism pathways and drug action pathways. PathWhiz is still being used to add to this very extensive collection (with the addition of about 2-3 pathways/week).

Overall, we believe that the graphical tools and visualization capabilities of today's modern web browsers have moved well-beyond the limitations associated with simplified wire-framed or schematic pathway diagrams. Indeed, many of these older pathway drawing tools and their associated pathway diagrams or databases were developed in the very earliest days of the internet. Given that more than a decade has passed since these early tools and earlier drawing conventions appeared,

we believe the time is ripe for a new generation of pathway drawing tools, a new generation of pathway databases and a new approach to generating, sharing and visualizing biological pathways on the internet. By releasing PathWhiz as a public web server and making the PathWhiz source code fully available, we are hoping that it will inspire others to contribute additional pathways to the scientific community and to further enhance both PathWhiz's capabilities and the quality of biological pathways found on the internet.

Chapter 5

Conclusions and Future Directions

5.1 Conclusions

This dissertation described the creation of the Yeast Metabolome Database (YMDB), the Small Molecule Pathway Database (SMPDB) and the PathWhiz pathway drawing system. The YMDB is a novel database that describes the metabolism and metabolome of *Saccharomyces cerevisiae* commonly known as baker's yeast. YMDB was inspired by the Human Metabolome Database (HMDB) and was constructed to provide a similar resource to those interested in yeast metabolism. The YMDB is the largest and most comprehensive compilation of yeast metabolism data and it is particularly rich in data regarding secondary yeast metabolites arising from fermentation processes on grape and barley substrates. YMDB has had over 75,000 unique visitors as of November 2013 since its initial release in January 2012. SMPDB (Version 2.0) was redesigned and rebuilt from the ground up to provide a more consistent, interactive and data driven resource for visualizing human metabolic, drug metabolism, drug action and disease pathways. Significant improvements were made in both the data quality and quantity as well as with SMPDB's visualization capabilities. As it currently stands, SMPDB is the most comprehensive small molecule pathway database ever created for human metabolism and among the largest pathway databases of any kind. To facilitate the redesign and reconstruction of SMPDB I developed the PathWhiz pathway drawing system. PathWhiz is a web based drawing tool that integrates data collection with pathway visualization to generate high quality, richly detailed and extensively hyperlinked pathway

drawings. PathWhiz appears to be the first web server ever developed that supports standardized pathway drawing.

Each of these systems or databases was developed to address a specific problem or a specific bottleneck in metabolomics. In particular, YMDB was developed to address the challenges of metabolite identification for yeast metabolomic studies and to facilitate metabolite studies with fermented foods and beverages (wine, beer, etc.). SMPDB was developed to simplify the biological interpretation of human metabolomic data and to increase our general understanding of human metabolism. To encourage data exchange and data standardization within the metabolomics community SMPDB and YMDB (as well as other databases I helped develop) were populated with spectral data in mzML and nmrML formats and pathway data in BioPAX format. Furthermore, YMDB has been formatted in a fully downloadable XML format. Finally, PathWhiz was developed to make biological pathway drawing easier and the formatting of biological pathways much more informative and far more standardized.

5.2 Future Directions

Much of the work I did over the past 3 years was aimed at trying to make metabolomics both easier and faster. In particular, the tools and computing infrastructure I developed for creating the YMDB were used again with the development of another microbial metabolomic database the E. coli Metabolome Database (ECMDB)[28]. These same concepts and the same software tools could potentially be recycled or revamped to construct other microbial metabolomic databases, such as those for Salmonella or Mycoplasma. Indeed, by generalizing some of the annotation features and database construction tools it may be possible to semi-automatically construct microbial metabolomic databases for hundreds of different microbes. This would be similar in concept to what has been done with KEGG and the Cyc databases [45, 7] but it would represent a significant advance since these metabolome databases would be much more comprehensive and far more useful for metabolomic research.

Similar ideas relating to the extension of pathway databases to other organisms

could also be applied using PathWhiz. For instance, the construction of the SMPDB was dependent on the existence of the HMDB. Without the metabolic information described with the HMDB it would not have been possible to create the SMPDB. With the construction of the YMDB it is now possible to use the PathWhiz system in conjunction with the YMDB to build a pathway database for yeast pathways. Furthermore the PathWhiz system could be utilized to generate metabolic pathways for almost any organism with well-defined metabolism. It might even be possible to use the human pathways provided by the SMPDB as a framework for creating these other organism specific databases more easily. This could be accomplished by comparing the metabolic processes found in human pathways with the desired organism while removing any non-complementary data. Further curation would obviously be required to correct for the differences and to fill out additional details. Additionally the PathWhiz system could be further developed to permit the facile generation of protein-protein interaction pathways as well as protein signaling pathways.

Bibliography

[1] ATANASSOV, I., HVARLEVA, T., RUSANOV, K., TSVETKOV, I., AND ATANASSOV, A. Wine Metabolite Profiling: Possible Application In Winemaking And Grapevine Breading In Bulgaria. *Biotechnology & Biotechnological Equipment 23*, 4 (2009), 1449–1452.

[2] BAIROCH, A. The Universal Protein Resource (UniProt). *Nucleic acids research 33*, Database issue (Dec. 2004), D154–D159.

[3] BARALDI, E., CARRARO, S., GIORDANO, G., AND RENIERO, F. Metabolomics: moving towards personalized medicine. *Italian Journal of Pediatrics 35*, 30 (2009).

[4] BOUHIFD, M., HARTUNG, T., HOGBERG, H. T., KLEENSANG, A., AND ZHAO, L. Review: toxicometabolomics. *Journal of applied toxicology : JAT 33*, 12 (Dec. 2013), 1365–1383.

[5] CASPI, R., ALTMAN, T., DALE, J. M., DREHER, K., FULCHER, C. A., GILHAM, F., KAIPA, P., KARTHIKEYAN, A. S., KOTHARI, A., KRUMMENACKER, M., LATENDRESSE, M., MUELLER, L. A., PALEY, S., POPESCU, L., PUJAR, A., SHEARER, A. G., ZHANG, P., AND KARP, P. D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research 38*, Database issue (Jan. 2010), D473–9.

[6] CASPI, R., ALTMAN, T., DREHER, K., FULCHER, C. A., SUBHRAVETI, P., KESELER, I. M., KOTHARI, A., KRUMMENACKER, M., LATENDRESSE, M., MUELLER, L. A., ONG, Q., PALEY, S., PUJAR, A., SHEARER, A. G., TRAVERS, M., WEERASINGHE, D., ZHANG, P., AND KARP, P. D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research 40*, Database issue (Jan. 2012), D742–53.

[7] CASPI, R., FOERSTER, H., FULCHER, C. A., HOPKINSON, R., INGRAHAM, J., KAIPA, P., KRUMMENACKER, M., PALEY, S., PICK, J., RHEE, S. Y., TISSIER, C., ZHANG, P., AND KARP, P. D. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic acids research 34*, Database issue (Jan. 2006), D511–6.

[8] CERAMI, E. G., GROSS, B. E., DEMIR, E., RODCHENKOV, I., BABUR, O., ANWAR, N., SCHULTZ, N., BADER, G. D., AND SANDER, C. Pathway Commons, a web resource for biological pathway data. *Nucleic acids research 39*, Database issue (Jan. 2011), D685–90.

[9] COSTANZO, M., BARYSHNIKOVA, A., BELLAY, J., KIM, Y., SPEAR, E. D., SEVIER, C. S., DING, H., KOH, J. L. Y., TOUFIGHI, K., MOSTAFAVI, S., PRINZ, J., ST ONGE, R. P., VANDERSLUIS, B., MAKHNEVYCH, T., VIZEACOUMAR, F. J., ALIZADEH, S., BAHR, S., BROST, R. L., CHEN, Y., COKOL, M., DESHPANDE, R., LI, Z., LIN, Z. Y., LIANG, W., MARBACK, M., PAW, J., SAN LUIS, B. J., SHUTERIQI, E., TONG, A. H. Y., VAN DYK, N., WALLACE, I. M., WHITNEY, J. A., WEIRAUCH, M. T., ZHONG, G., ZHU, H., HOURY, W. A., BRUDNO, M., RAGIBIZADEH, S., PAPP, B., PAL, C., ROTH, F. P., GIAEVER, G., NISLOW, C., TROYANSKAYA, O. G., BUSSEY, H., BADER, G. D., GINGRAS, A. C., MORRIS, Q. D., KIM, P. M., KAISER, C. A., MYERS, C. L., ANDREWS, B. J., AND BOONE, C. The Genetic Landscape of a Cell. *Science 327*, 5964 (Jan. 2010), 425–431.

[10] COSTANZO, M., AND BOONE, C. SGAM: an array-based approach for high-resolution genetic mapping in Saccharomyces cerevisiae. *Methods in molecular biology (Clifton, N.J.) 548* (2009), 37–53.

[11] CROFT, D., O'KELLY, G., WU, G., HAW, R., GILLESPIE, M., MATTHEWS, L., CAUDY, M., GARAPATI, P., GOPINATH, G., JASSAL, B., JUPE, S., KALATSKAYA, I., MAHAJAN, S., MAY, B., NDEGWA, N., SCHMIDT, E., SHAMOVSKY, V., YUNG, C., BIRNEY, E., HERMJAKOB, H., D'EUSTACHIO, P., AND STEIN, L. Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research 39*, Database issue (Jan. 2011), D691–7.

[12] DE MATOS, P., ALCÁNTARA, R., DEKKER, A., ENNIS, M., HASTINGS, J., HAUG, K., SPITERI, I., TURNER, S., AND STEINBECK, C. Chemical Entities of Biological Interest: an update. *Nucleic acids research 38*, Database issue (Jan. 2010), D249–54.

[13] DEGTYARENKO, K., DE MATOS, P., ENNIS, M., HASTINGS, J., ZBINDEN, M., MCNAUGHT, A., ALCÁNTARA, R., DARSOW, M., GUEDJ, M., AND ASHBURNER, M. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research 36*, Database issue (Jan. 2008), D344–50.

[14] DEMIR, E., CARY, M. P., PALEY, S., FUKUDA, K., LEMER, C., VASTRIK, I., WU, G., D'EUSTACHIO, P., SCHAEFER, C., LUCIANO, J., SCHACHERER, F., MARTINEZ-FLORES, I., HU, Z., JIMENEZ-JACINTO, V., JOSHI-TOPE, G., KANDASAMY, K., LOPEZ-FUENTES, A. C., MI, H., PICHLER, E., RODCHENKOV, I., SPLENDIANI, A., TKACHEV, S., ZUCKER, J., GOPINATH, G., RAJASIMHA, H., RAMAKRISHNAN, R., SHAH, I., SYED, M., ANWAR, N., BABUR, O., BLINOV, M., BRAUNER, E., CORWIN, D., DONALDSON, S., GIBBONS, F., GOLDBERG, R., HORNBECK, P., LUNA, A., MURRAY-RUST, P., NEUMANN, E., REUBENACKER, O., SAMWALD, M., VAN IERSEL, M., WIMALARATNE, S., ALLEN, K., BRAUN, B., WHIRL-CARRILLO, M., CHEUNG, K.-H., DAHLQUIST, K., FINNEY, A., GILLESPIE, M., GLASS, E., GONG, L., HAW, R., HONIG, M., HUBAUT, O., KANE, D., KRUPA, S., KUTMON, M., LEONARD, J., MARKS, D., MERBERG, D., PETRI, V., PICO, A., RAVENSCROFT, D., REN, L., SHAH, N., SUNSHINE, M., TANG, R., WHALEY, R., LETOVKSY, S., BUETOW, K. H., RZHETSKY, A., SCHACHTER, V., SOBRAL, B. S., DOGRUSOZ, U., MCWEENEY, S., ALADJEM, M., BIRNEY, E., COLLADO-VIDES, J., GOTO, S., HUCKA, M., NOVÈRE, N. L., MALTSEV, N., PANDEY, A., THOMAS, P., WINGENDER, E., KARP, P. D., SANDER, C.,

AND BADER, G. D. The BioPAX community standard for pathway data sharing. *Nature biotechnology 28*, 9 (Sept. 2010), 935–942.

[15] DÉNES, J., SZABÓ, E., ROBINETTE, S. L., SZATMÁRI, I., SZŐNYI, L., KREUDER, J. G., RAUTERBERG, E. W., AND TAKÁTS, Z. Metabonomics of newborn screening dried blood spot samples: a novel approach in the screening and diagnostics of inborn errors of metabolism. *Analytical chemistry 84*, 22 (Nov. 2012), 10113–10120.

[16] DWORZANSKI, J. P., SNYDER, A. P., CHEN, R., ZHANG, H., WISHART, D., AND LI, L. Identification of bacteria using tandem mass spectrometry combined with a proteome database and statistical scoring. *Analytical chemistry 76*, 8 (Apr. 2004), 2355–2366.

[17] ELLIOTT, B., KIRAC, M., CAKMAK, A., YAVAS, G., MAYES, S., CHENG, E., WANG, Y., GUPTA, C., OZSOYOGLU, G., AND MERAL OZSOYOGLU, Z. PathCase: pathways database system. *Bioinformatics 24*, 21 (Oct. 2008), 2526–2533.

[18] ENGEL, S. R., BALAKRISHNAN, R., BINKLEY, G., CHRISTIE, K. R., COSTANZO, M. C., DWIGHT, S. S., FISK, D. G., HIRSCHMAN, J. E., HITZ, B. C., HONG, E. L., KRIEGER, C. J., LIVSTONE, M. S., MIYASATO, S. R., NASH, R., OUGHTRED, R., PARK, J., SKRZYPEK, M. S., WENG, S., WONG, E. D., DOLINSKI, K., BOTSTEIN, D., AND CHERRY, J. M. Saccharomyces Genome Database provides mutant phenotype data. *Nucleic acids research 38*, Database issue (Jan. 2010), D433–6.

[19] FIEHN, O. Metabolomics – the link between genotypes and phenotypes. *Plant Molecular Biology 48*, 1/2 (2002), 155–171.

[20] FROLKIS, A., KNOX, C., LIM, E., JEWISON, T., LAW, V., HAU, D. D., LIU, P., GAUTAM, B., LY, S., GUO, A. C., XIA, J., LIANG, Y., SHRIVASTAVA, S., AND WISHART, D. S. SMPDB: The Small Molecule Pathway Database. *Nucleic acids research 38*, Database issue (Jan. 2010), D480–7.

[21] FUNAHASHI, A., MOROHASHI, M., AND KITANO, H. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO 1*, 5 (Nov. 2003), 159–162.

[22] GERMAN, J. B., BAUMAN, D. E., BURRIN, D. G., FAILLA, M. L., FREAKE, H. C., KING, J. C., KLEIN, S., MILNER, J. A., PELTO, G. H., RASMUSSEN, K. M., AND ZEISEL, S. H. Metabolomics in the Opening Decade of the 21st Century: Building the Roads to Individualized Health. *The Journal of nutrition 134* (2004), 2729–2732.

[23] GERMAN, J. B., HAMMOCK, B. D., AND WATKINS, S. M. Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics 1*, 1 (Mar. 2005), 3–9.

[24] GERMAN, J. B., ROBERTS, M.-A., AND WATKINS, S. M. Genomics and Metabolomics as Markers for the Interaction of Diet and Health: Lessons from Lipids. *The Journal of nutrition* (2003).

[25] GO, E. P. Database resources in metabolomics: an overview. *Journal of neuroimmune pharmacology : the official journal of the Society on NeuroImmune Pharmacology 5*, 1 (Mar. 2010), 18–30.

[26] GOFFEAU, A., BARRELL, B. G., BUSSEY, H., DAVIS, R. W., DUJON, B., FELDMANN, H., GALIBERT, F., HOHEISEL, J. D., JACQ, C., JOHNSTON, M., LOUIS, E. J., MEWES, H. W., MURAKAMI, Y., PHILIPPSEN, P., TETTELIN, H., AND OLIVER, S. G. Life with 6000 Genes. *Science 274*, 5287 (Oct. 1996), 546–567.

[27] GÜLDENER, U., MÜNSTERKÖTTER, M., KASTENMÜLLER, G., STRACK, N., VAN HELDEN, J., LEMER, C., RICHELLES, J., WODAK, S. J., GARCÍA-MARTÍNEZ, J., PÉREZ-ORTÍN, J. E., MICHAEL, H., KAPS, A., TALLA, E., DUJON, B., ANDRÉ, B., SOUCIET, J. L., DE MONTIGNY, J., BON, E., GAILLARDIN, C., AND MEWES, H. W. CYGD: the Comprehensive Yeast Genome Database. *Nucleic acids research 33*, Database issue (Jan. 2005), D364–8.

[28] GUO, A. C., JEWISON, T., WILSON, M., LIU, Y., KNOX, C., DJOUMBOU, Y., LO, P., MANDAL, R., KRISHNAMURTHY, R., AND WISHART, D. S. ECMDB: the E. coli Metabolome Database. *Nucleic acids research 41*, Database issue (Jan. 2013), D625–30.

[29] HAMOSH, A., SCOTT, A. F., AMBERGER, J. S., BOCCHINI, C. A., AND MCKUSICK, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research 33*, Database issue (2005), D514–7.

[30] HASTINGS, J., DE MATOS, P., DEKKER, A., ENNIS, M., HARSHA, B., KALE, N., MUTHUKRISHNAN, V., OWEN, G., TURNER, S., WILLIAMS, M., AND STEINBECK, C. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic acids research 41*, Database issue (Jan. 2013), D456–63.

[31] HELLER, M., VITALI, L., OLIVEIRA, M. A. L., COSTA, A. C. O., AND MICKE, G. A. A rapid sample screening method for authenticity control of whiskey using capillary electrophoresis with online preconcentration. *Journal of agricultural and food chemistry 59*, 13 (July 2011), 6882–6888.

[32] HERRGÅRD, M. J., SWAINSTON, N., DOBSON, P., DUNN, W. B., ARGA, K. Y., ARVAS, M., BLÜTHGEN, N., BORGER, S., COSTENOBLE, R., HEINEMANN, M., HUCKA, M., LE NOVÈRE, N., LI, P., LIEBERMEISTER, W., MO, M. L., OLIVEIRA, A. P., PETRANOVIC, D., PETTIFER, S., SIMEONIDIS, E., SMALLBONE, K., SPASIĆ, I., WEICHART, D., BRENT, R., BROOMHEAD, D. S., WESTERHOFF, H. V., KIRDAR, B., PENTTILÄ, M., KLIPP, E., PALSSON, B. Ø., SAUER, U., OLIVER, S. G., MENDES, P., NIELSEN, J., AND KELL, D. B. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature biotechnology 26*, 10 (Oct. 2008), 1155–1160.

[33] HODGES, P. E., MCKEE, A. H., DAVIS, B. P., PAYNE, W. E., AND GARRELS, J. I. The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic acids research 27*, 1 (Jan. 1999), 69–73.

[34] HONG, Y.-S. NMR-based metabolomics in wine science. *Magnetic resonance in chemistry : MRC 49 Suppl 1* (Dec. 2011), S13–21.

[35] HORAI, H., ARITA, M., KANAYA, S., NIHEI, Y., IKEDA, T., SUWA, K., OJIMA, Y., TANAKA, K., TANAKA, S., AOSHIMA, K., ODA, Y., KAKAZU, Y., KUSANO, M., TOHGE, T., MATSUDA, F., SAWADA, Y., HIRAI, M. Y., NAKANISHI, H., IKEDA, K., AKIMOTO, N., MAOKA, T., TAKAHASHI, H., ARA, T., SAKURAI, N., SUZUKI, H., SHIBATA, D., NEUMANN, S., IIDA, T., TANAKA, K., FUNATSU, K., MATSUURA, F., SOGA, T., TAGUCHI, R., SAITO, K., AND NISHIOKA, T. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry : JMS 45*, 7 (July 2010), 703–714.

[36] HU, Z., NG, D. M., YAMADA, T., CHEN, C., KAWASHIMA, S., MELLOR, J., LINGHU, B., KANEHISA, M., STUART, J. M., AND DELISI, C. VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. *Nucleic acids research 35*, Web Server (May 2007), W625–W632.

[37] INUI, T., TSUCHIYA, F., ISHIMARU, M., OKA, K., AND KOMURA, H. Different beers with different hops. Relevant compounds for their aroma characteristics. *Journal of agricultural and food chemistry 61*, 20 (May 2013), 4758–4764.

[38] JANEČKOVÁ, H., HRON, K., WOJTOWICZ, P., HLÍDKOVÁ, E., BAREŠOVÁ, A., FRIEDECKÝ, D., ZÍDKOVÁ, L., HORNIK, P., BEHÚLOVÁ, D., PROCHÁZKOVÁ, D., VINOHRADSKÁ, H., PEŠKOVÁ, K., BRUHEIM, P., SMOLKA, V., SŤASTNÁ, S., AND ADAM, T. Targeted metabolomic analysis of plasma samples for the diagnosis of inherited metabolic disorders. *Journal of chromatography. A 1226* (Feb. 2012), 11–17.

[39] JERNIGAN, D. H. The global alcohol industry: an overview. *Addiction (Abingdon, England) 104 Suppl 1* (Feb. 2009), 6–12.

[40] JEWISON, T., KNOX, C., NEVEU, V., DJOUMBOU, Y., GUO, A. C., LEE, J., LIU, P., MANDAL, R., KRISHNAMURTHY, R., SINELNIKOV, I., WILSON, M., AND WISHART, D. S. YMDB: the Yeast Metabolome Database. *Nucleic acids research 40*, Database issue (Jan. 2012), D815–20.

[41] JEWISON, T., SU, Y., DISFANY, F. M., LIANG, Y., KNOX, C., MACIEJEWSKI, A., POELZER, J., HUYNH, J., ZHOU, Y., ARNDT, D., DJOUMBOU, Y., LIU, Y., DENG, L., GUO, A. C., HAN, B., PON, A., WILSON, M., RAFATNIA, S., LIU, P., AND WISHART, D. S. SMPDB 2.0: Big Improvements to the Small Molecule Pathway Database. *Nucleic acids research* (Nov. 2013).

[42] JOSHI-TOPE, G., GILLESPIE, M., VASTRIK, I., D'EUSTACHIO, P., SCHMIDT, E., DE BONO, B., JASSAL, B., GOPINATH, G. R., WU, G. R., MATTHEWS, L., LEWIS, S., BIRNEY, E., AND STEIN, L. Reactome: a knowledgebase of biological pathways. *Nucleic acids research 33*, Database issue (Jan. 2005), D428–32.

[43] KANEHISA, M., GOTO, S., FURUMICHI, M., TANABE, M., AND HIRAKAWA, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research 38*, Database issue (Jan. 2010), D355–60.

[44] KANEHISA, M., GOTO, S., KAWASHIMA, S., AND NAKAYA, A. The KEGG databases at GenomeNet. *Nucleic acids research 30*, 1 (2002), 42–46.

[45] KANEHISA, M., GOTO, S., SATO, Y., FURUMICHI, M., AND TANABE, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research 40*, Database issue (Jan. 2012), D109–14.

[46] KARP, P., RILEY, M., AND PALEY, S. The MetaCyc Database. *Nucleic acids research 30*, 1 (2002), 59–61.

[47] KARP, P. D., PALEY, S. M., KRUMMENACKER, M., LATENDRESSE, M., DALE, J. M., LEE, T. J., KAIPA, P., GILHAM, F., SPAULDING, A., POPESCU, L., ALTMAN, T., PAULSEN, I., KESELER, I. M., AND CASPI, R. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics 11*, 1 (Jan. 2010), 40–79.

[48] KELDER, T., VAN IERSEL, M. P., HANSPERS, K., KUTMON, M., CONKLIN, B. R., EVELO, C. T., AND PICO, A. R. WikiPathways: building research communities on biological pathways. *Nucleic acids research 40*, Database issue (Dec. 2011), D1301–D1307.

[49] KESELER, I. M., MACKIE, A., PERALTA-GIL, M., SANTOS-ZAVALETA, A., GAMA-CASTRO, S., BONAVIDES-MARTÍNEZ, C., FULCHER, C., HUERTA, A. M., KOTHARI, A., KRUMMENACKER, M., LATENDRESSE, M., MUÑIZ-RASCADO, L., ONG, Q., PALEY, S., SCHRÖDER, I., SHEARER, A. G., SUBHRAVETI, P., TRAVERS, M., WEERASINGHE, D., WEISS, V., COLLADO-VIDES, J., GUNSALUS, R. P., PAULSEN, I., AND KARP, P. D. EcoCyc: fusing model organism databases with systems biology. *Nucleic acids research 41*, Database issue (Jan. 2013), D605–12.

[50] KIND, T., TOLSTIKOV, V., FIEHN, O., AND WEISS, R. H. A comprehensive urinary metabolomic approach for identifying kidney cancer. *Analytical biochemistry 363*, 2 (Apr. 2007), 185–195.

[51] KNOX, C., LAW, V., JEWISON, T., LIU, P., LY, S., FROLKIS, A., PON, A., BANCO, K., MAK, C., NEVEU, V., DJOUMBOU, Y., EISNER, R., GUO, A. C., AND WISHART, D. S. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research 39*, Database issue (Jan. 2011), D1035–41.

[52] KNOX, C., SHRIVASTAVA, S., STOTHARD, P., EISNER, R., AND WISHART, D. S. BioSpider: a web server for automating metabolome annotations. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2007), 145–156.

[53] KOPKA, J., SCHAUER, N., KRUEGER, S., BIRKEMEYER, C., USADEL, B., BERGMÜLLER, E., DÖRMANN, P., WECKWERTH, W., GIBON, Y., STITT, M., WILLMITZER, L., FERNIE, A. R., AND STEINHAUSER, D. GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics 21*, 8 (Apr. 2005), 1635–1638.

[54] LIM, E., PON, A., DJOUMBOU, Y., KNOX, C., SHRIVASTAVA, S., GUO, A. C., NEVEU, V., AND WISHART, D. S. T3DB: a comprehensively annotated database of common toxins and their targets. *Nucleic acids research 38*, Database issue (Jan. 2010), D781–6.

[55] LINSTROM, P. J., AND MALLARD, W. G. The NIST Chemistry WebBook: A Chemical Data Resource on the Internet . *Journal of Chemical & Engineering Data 46*, 5 (Sept. 2001), 1059–1063.

[56] MARTENS, L., CHAMBERS, M., STURM, M., KESSNER, D., LEVANDER, F., SHOFSTAHL, J., TANG, W. H., ROMPP, A., NEUMANN, S., PIZARRO, A. D., MONTECCHI-PALAZZI, L., TASMAN, N., COLEMAN, M., REISINGER, F., SOUDA, P., HERMJAKOB, H., BINZ, P. A., AND DEUTSCH, E. W. mzML–a Community Standard for Mass Spectrometry Data. *Molecular & Cellular Proteomics 10*, 1 (Dec. 2010), R110.000133–R110.000133.

[57] MATTHEWS, L., GOPINATH, G., GILLESPIE, M., CAUDY, M., CROFT, D., DE BONO, B., GARAPATI, P., HEMISH, J., HERMJAKOB, H., JASSAL, B., KANAPIN, A., LEWIS, S., MAHAJAN, S., MAY, B., SCHMIDT, E., VASTRIK, I., WU, G., BIRNEY, E., STEIN, L., AND D'EUSTACHIO, P. Reactome knowledgebase of human biological pathways and processes. *Nucleic acids research 37*, Database issue (Jan. 2009), D619–22.

[58] MICHAL, G. On representation of metabolic pathways. *Bio Systems 47*, 1-2 (June 1998), 1–7.

[59] MOCO, S., VERVOORT, J., MOCO, S., BINO, R. J., DE VOS, R. C. H., AND BINO, R. Metabolomics technologies and metabolite identification. *TrAC Trends in Analytical Chemistry 26*, 9 (Oct. 2007), 855–866.

[60] MURRAY-RUST, P., AND RZEPA, H. S. Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles. *Journal of Chemical Information and Modeling 39*, 6 (Nov. 1999), 928–942.

[61] PEDRO, M. Emerging bioinformatics for the metabolome. *Briefings in Bioinformatics 3*, 2 (Jan. 2002), 134–145.

[62] PENCE, H. E., AND WILLIAMS, A. ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education 87*, 11 (Nov. 2010), 1123–1124.

[63] PICO, A. R., KELDER, T., VAN IERSEL, M. P., HANSPERS, K., CONKLIN, B. R., AND EVELO, C. WikiPathways: pathway editing for the people. *PLoS biology 6*, 7 (July 2008), e184.

[64] PUTRI, S. P., NAKAYAMA, Y., MATSUDA, F., UCHIKATA, T., KOBAYASHI, S., MATSUBARA, A., AND FUKUSAKI, E. Current metabolomics: Practical applications. *Journal of bioscience and bioengineering 115*, 6 (June 2013), 579–589.

[65] ROBINSON, M. D., GRIGULL, J., MOHAMMAD, N., AND HUGHES, T. R. FunSpec: a web-based cluster interpreter for yeast. *BMC bioinformatics 3* (Nov. 2002), 35.

[66] SCALBERT, A., BRENNAN, L., FIEHN, O., HANKEMEIER, T., KRISTAL, B. S., VAN OMMEN, B., PUJOS-GUILLOT, E., VERHEIJ, E., WISHART, D., AND WOPEREIS, S. Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics 5*, 4 (Dec. 2009), 435–458.

[67] SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N. S., WANG, J. T., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B., AND IDEKER, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research 13*, 11 (Nov. 2003), 2498–2504.

[68] SINGH, O. V. Proteomics and metabolomics: The molecular make-up of toxic aromatic pollutant bioremediation. *PROTEOMICS 6*, 20 (Oct. 2006), 5481–5492.

[69] SLUPSKY, C. M., RANKIN, K. N., FU, H., CHANG, D., ROWE, B. H., CHARLES, P. G. P., MCGEER, A., LOW, D., LONG, R., KUNIMOTO, D., SAWYER, M. B., FEDORAK, R. N., ADAMKO, D. J., SAUDE, E. J., SHAH, S. L., AND MARRIE, T. J. Pneumococcal Pneumonia: Potential for Diagnosis through a Urinary Metabolic Profile. *Journal of Proteome Research 8*, 12 (Dec. 2009), 5550–5558.

[70] STEINBECK, C., KRAUSE, S., AND KUHN, S. NMRShiftDB-constructing a free chemical information system with open-source components. *Journal of Chemical Information and Computer Sciences 43*, 6 (Nov. 2003), 1733–1739.

[71] STRÖMBÄCK, L., AND LAMBRIX, P. Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics 21*, 24 (Dec. 2005), 4401–4407.

[72] SUNDARARAJ, S., GUO, A., NAZHAD, B. H., ROUANI, M., STOTHARD, P., ELLISON, M., AND WISHART, D. S. The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of Escherichia coli. *Nucleic acids research 32*, 90001 (Jan. 2004), 293D–295.

[73] THORN, C. F., KLEIN, T. E., AND ALTMAN, R. B. PharmGKB: The Pharmacogenomics Knowledge Base. In *Pharmacogenomics*. Humana Press, Totowa, NJ, May 2013, pp. 311–320.

[74] ULRICH, E. L., AKUTSU, H., DORELEIJERS, J. F., HARANO, Y., IOANNIDIS, Y. E., LIN, J., LIVNY, M., MADING, S., MAZIUK, D., MILLER, Z., NAKATANI, E., SCHULTE, C. F., TOLMIE, D. E., KENT WENGER, R., YAO, H., AND MARKLEY, J. L. BioMagResBank. *Nucleic acids research 36*, Database (Dec. 2007), D402–D408.

[75] UNIPROT CONSORTIUM. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic acids research 41*, Database issue (Jan. 2013), D43–7.

[76] VAN DER GREEF, J., HANKEMEIER, T., AND MCBURNEY, R. N. Metabolomics-based systems biology and personalized medicine: moving towards n = 1 clinical trials? *Pharmacogenomics 7*, 7 (Oct. 2006), 1087–1094.

[77] VAN IERSEL, M. P., KELDER, T., PICO, A. R., HANSPERS, K., COORT, S., CONKLIN, B. R., AND EVELO, C. Presenting and exploring biological pathways with PathVisio. *BMC bioinformatics 9* (2008), 399.

[78] WANG, Y., XIAO, J., SUZEK, T. O., ZHANG, J., WANG, J., AND BRYANT, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research 37*, Web Server issue (2009), W623–3.

[79] WEBB, R. L., AND MA'AYAN, A. Sig2BioPAX: Java tool for converting flat files to BioPAX Level 3 format. *Source code for biology and medicine 6* (2011), 5.

[80] WECKWERTH, W. Metabolomics: an integral technique in systems biology. *Bioanalysis 2*, 4 (Apr. 2010), 829–836.

[81] WEININGER, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences 28* (1988), 31–36.

[82] WIKOFF, W. R., GANGOITI, J. A., BARSHOP, B. A., AND SIUZDAK, G. Metabolomics identifies perturbations in human disorders of propionate metabolism. *Clinical chemistry 53*, 12 (Dec. 2007), 2169–2176.

[83] WISHART, D. S. Current progress in computational metabolomics. *Briefings in Bioinformatics 8*, 5 (Sept. 2007), 279–293.

[84] WISHART, D. S. Applications of Metabolomics in Drug Discovery and Development. *Drugs in R & D 9*, 5 (2008), 307–322.

[85] WISHART, D. S. Metabolomics: applications to food science and nutrition research. *Trends in Food Science & Technology 19*, 9 (Sept. 2008), 482–493.

[86] WISHART, D. S., JEWISON, T., GUO, A. C., WILSON, M., KNOX, C., LIU, Y., DJOUMBOU, Y., MANDAL, R., AZIAT, F., DONG, E., BOUATRA, S., SINELNIKOV, I., ARNDT, D., XIA, J., LIU, P., YALLOU, F., BJORNDAHL, T., PEREZ-PINEIRO, R., EISNER, R., ALLEN, F., NEVEU, V., GREINER, R., AND SCALBERT, A. HMDB 3.0–The Human Metabolome Database in 2013. *Nucleic acids research 41*, Database issue (Jan. 2013), D801–7.

[87] WISHART, D. S., KNOX, C., GUO, A. C., EISNER, R., YOUNG, N., GAUTAM, B., HAU, D. D., PSYCHOGIOS, N., DONG, E., BOUATRA, S., MANDAL, R., SINELNIKOV, I., XIA, J., JIA, L., CRUZ, J. A., LIM, E., SOBSEY, C. A., SHRIVASTAVA, S., HUANG, P., LIU, P., FANG, L., PENG, J., FRADETTE, R., CHENG, D., TZUR, D., CLEMENTS, M., LEWIS, A., DE SOUZA, A., ZUNIGA, A., DAWE, M., XIONG, Y., CLIVE, D., GREINER, R., NAZYROVA, A., SHAYKHUTDINOV, R., LI, L., VOGEL, H. J., AND FORSYTHE, I. HMDB: a knowledgebase for the human metabolome. *Nucleic acids research 37*, Database issue (Jan. 2009), D603–10.

[88] WISHART, D. S., KNOX, C., GUO, A. C., SHRIVASTAVA, S., HASSANALI, M., STOTHARD, P., CHANG, Z., AND WOOLSEY, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research 34*, Database issue (Jan. 2006), D668–72.

[89] WISHART, D. S., TZUR, D., KNOX, C., EISNER, R., GUO, A. C., YOUNG, N., CHENG, D., JEWELL, K., ARNDT, D., SAWHNEY, S., FUNG, C., NIKOLAI, L., LEWIS, M., COUTOULY, M.-A., FORSYTHE, I., TANG, P., SHRIVASTAVA, S., JERONCIC, K., STOTHARD, P., AMEGBEY, G., BLOCK, D., HAU, D. D., WAGNER, J., MINIACI, J., CLEMENTS, M., GEBREMEDHIN, M., GUO, N., ZHANG, Y., DUGGAN, G. E., MACINNIS, G. D., WELJIE, A. M., DOWLATABADI, R., BAMFORTH, F., CLIVE, D., GREINER, R., LI, L., MARRIE, T., SYKES, B. D., VOGEL, H. J., AND QUERENGESSER, L. HMDB: the Human Metabolome Database. *Nucleic acids research 35*, Database issue (Jan. 2007), D521–6.

[90] WOHLGEMUTH, G., HALDIYA, P. K., WILLIGHAGEN, E., KIND, T., AND FIEHN, O. The Chemical Translation Service—a web-based tool to improve standardization of metabolomic reports. *Bioinformatics 26*, 20 (2010), 2647–2648.

[91] XIA, J., BJORNDAHL, T. C., TANG, P., AND WISHART, D. S. MetaboMiner–semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC bioinformatics 9* (2008), 507.

[92] XIA, J., PSYCHOGIOS, N., YOUNG, N., AND WISHART, D. S. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic acids research 37*, Web Server issue (July 2009), W652–60.

[93] XIA, J., AND WISHART, D. S. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic acids research 38*, Web Server issue (July 2010), W71–7.