D480–D487 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 30 November 2009*
doi:10.1093/nar/gkp1002

SMPDB: The Small Molecule Pathway Database

Alex Frolkis[1], Craig Knox[1], Emilia Lim[1], Timothy Jewison[1], Vivian Law[2], David D. Hau[1], Phillip Liu[3], Bijaya Gautam[1], Son Ly[2], An Chi Guo[1], Jianguo Xia[3], Yongjie Liang[1], Savita Shrivastava[1] and David S. Wishart[1,2,3,4,*]

[1]Department of Computing Science, University of Alberta, Edmonton, AB, Canada T6G 2E8, [2]Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, AB, Canada T6G 2N8, [3]Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada T6G 2E8 and [4]National Institute for Nanotechnology, 11421 Saskatchewan Drive, Edmonton, AB, Canada T6G 2M9

Received August 15, 2009; Revised October 14, 2009; Accepted October 18, 2009

ABSTRACT

The Small Molecule Pathway Database (SMPDB) is an interactive, visual database containing more than 350 small-molecule pathways found in humans. More than 2/3 of these pathways (>280) are not found in any other pathway database. SMPDB is designed specifically to support pathway elucidation and pathway discovery in clinical metabolomics, transcriptomics, proteomics and systems biology. SMPDB provides exquisitely detailed, hyperlinked diagrams of human metabolic pathways, metabolic disease pathways, metabolite signaling pathways and drug-action pathways. All SMPDB pathways include information on the relevant organs, organelles, subcellular compartments, protein cofactors, protein locations, metabolite locations, chemical structures and protein quaternary structures. Each small molecule is hyperlinked to detailed descriptions contained in the Human Metabolome Database (HMDB) or DrugBank and each protein or enzyme complex is hyperlinked to UniProt. All SMPDB pathways are accompanied with detailed descriptions, providing an overview of the pathway, condition or processes depicted in each diagram. The database is easily browsed and supports full text searching. Users may query SMPDB with lists of metabolite names, drug names, genes/protein names, SwissProt IDs, GenBank IDs, Affymetrix IDs or Agilent microarray IDs. These queries will produce lists of matching pathways and highlight the matching molecules on each of the pathway diagrams. Gene, metabolite and protein concentration data can also be visualized through SMPDB's mapping interface. All of SMPDB's images, image maps, descriptions and tables are downloadable. SMPDB is available at: http://www.smpdb.ca.

INTRODUCTION

Much of the analytical technology developed over the past two decades to facilitate genomics, transcriptomics, proteomics and metabolomics has been aimed at helping to elucidate the connections between genes, proteins and metabolites and their physiological consequences. These connections, their processes and outcomes can usually be summarized in pathway diagrams or pathway maps. Indeed, pathway maps in textbooks and wall charts have long been used to describe the basics of metabolism. They have also proven to be valuable tools in giving some rhyme and reason to complex and interconnected metabolic processes (1). The concept of metabolic pathway mapping has actually proven to be so effective that it has been extended to describe protein signaling, protein–DNA interactions and many other molecular biological phenomena.

Perhaps the first molecular pathway resource to gain wide popularity was a paper wall chart developed by Dr. Gerhard Michal, a staff scientist at Boehringer Mannheim (now Roche Applied Sciences), and published in 1968 (2). This richly annotated and compactly illustrated poster depicts most of the known metabolic pathways, metabolites and enzymes associated primarily with eukaryotic metabolism. It has been revised and expanded many times since its first release and has proven to be very popular, with more than 1 million copies now in print. With the development of the World Wide Web along the growing interest in systems biology (3) and model organism pathways (4), the utility of wall charts has diminished somewhat and a substantial shift has been made to displaying this information through web-accessible pathway charts and pathway databases. Over the past decade, a number of superb, web-accessible

*To whom correspondence should be addressed. Tel: +1 780 492 0383; Fax: +1 780 492 1071; Email: david.wishart@ualberta.ca

molecular pathway resources have emerged, including Kyoto Encyclopedia of Genes and Genomes (KEGG) (5), the 'Cyc' databases (6,7), the Reactome database (8), WikiPathways/GeneMAPP (9), the Edinburgh Human Metabolic Network (EHMN) (10), PharmGKB (11), the Medical Biochemistry Page (http://themedicalbiochemistrypage.org/) and PathwayCommons (http://www.pathwaycommons.org). Several companies have also created beautifully illustrated, freely available pathway databases such as BioCarta (www.biocarta.com), Cell Signal pathways (www.cellsignal.com), Sigma-Aldrich pathways (http://www.sigmaaldrich.com/life-science/cell-biology/learning-center/) Ambion pathways (http://www.ambion.com), Calbiochem/Merck pathways (http://www.merckbiosciences.co.uk) and ProteinLounge (http://www.proteinlounge.com/). A number of commercial pathway databases also exist, such as TransPath (from BioBase Inc.), PathArt (from Jubilant Biosys Inc.), MetaBase (from GeneGo Inc.) and Ingenuity Pathways Analysis (Ingenuity Systems Inc.), many of which provide impressive numbers of molecular pathway diagrams.

Almost all of the above-mentioned databases contain a mix of metabolic, cell signaling and disease pathways. A few, such as KEGG, the 'Cyc' databases and Reactome, have a much greater emphasis on small-molecule interactions (i.e. metabolism) while most of the company-generated databases place a greater emphasis on protein–protein interactions (i.e. signaling pathways). Most pathway databases maintain data on a large number of different organisms (microbes to plants to animals), while a small number are very organism specific (The Medical Biochemistry Page, EcoCyc). Some pathway resources such as KEGG and Reactome use pathway diagrams that are very generic and highly schematic or simplified, while still others use diagrams that very specific and rich in detail, color and content (BioCarta, ProteinLounge, Ambion). Most of these web-based databases support interactive image mapping with hyperlinked information content that can range from very sparse (most of the open-access company databases) to moderately detailed (KEGG, Reactome, the 'Cyc' databases). Almost all of the databases support some kind of limited text search and a few, such as Reactome, Pathway Commons and the 'Cyc' databases, support the mapping of gene, protein and/or metabolite expression data onto pathway diagrams. Overall, the quality, coverage and capability of today's molecular pathway databases are quite impressive and they are frequently used in studies relating to proteomics, systems biology, metabolic engineering and metabolomics. However, all is not well in the pathway database field.

Over the past few years, our group has been actively involved in clinical metabolomics (12,13) and in developing a number of clinical metabolomics databases, including DrugBank (14), the HMDB (15) and the T3DB (http://www.t3db.org). In the course of this work, we have used or become very familiar with many of the pathway resources described above. Unfortunately, with respect to clinical metabolomics (i.e. human metabolism, human diseases and drug therapy), we have found that none of the pathway databases mentioned above has the depth or breadth of information needed to be truly effective in this particular area of research. This situation is certainly understandable given that most pathway databases were developed before the advent of metabolomics or without the intent of creating a resource specifically for clinical metabolomics.

Ideally, what is needed for clinical metabolomics is a pathway database that focuses on small molecules only and which provides large numbers (>300) of detailed, hyperlinked diagrams of human metabolic pathways, metabolic disease pathways, metabolite signaling pathways and drug-action or drug-metabolism pathways. To be truly useful, these pathway diagrams need to include information on the relevant organs, organelles, subcellular compartments, protein cofactors, protein locations, metabolite locations, chemical structures and protein quaternary structures. Each small molecule, gene or protein shown in each pathway diagram should hyperlinked to detailed descriptions as well as relevant clinical information and all pathways should be accompanied with overviews describing the corresponding pathway, condition or processes. The database should be easily browsed and must support full text, structure and sequence searching. Likewise, gene, metabolite and protein lists or concentration data should be easily mapped to every one of the pathway diagrams. In response to these very specific needs, we have developed just such a database: the Small Molecule Pathway Database (SMPDB).

DATABASE DESCRIPTION

SMPDB is a pathway database designed to facilitate clinical 'omics' studies, with a specific emphasis on clinical metabolomics. In particular, SMPD consists of more than 350 hand-drawn pathways describing small-molecule metabolism or small-molecule processes that are specific to humans. More than 280 of these are unique to SMPDB and are found nowhere else. These pathways fall into four different categories: (i) metabolic pathways; (ii) small-molecule disease pathways; (iii) small-molecule drug pathways; and (iv) small-molecule signaling pathways. In determining whether a small-molecule pathway or process is suitable for inclusion in SMPDB, the pathway must be found in humans and it must have at least five small molecules (if it is a metabolic pathway) or its central feature must be based on the action of at least one small molecule (if it is a disease, drug or signaling pathway).

More specifically, in SMPDB, disease pathways refer to those pathways describing human disease processes (cause and effect) where small-molecule metabolite dysregulation is the primary hallmark of the disease. That is, if significant concentration changes for a small molecule or set of small molecules is commonly used for the diagnosis, prognosis or monitoring of a given disease, then it qualifies for inclusion in SMPDB. For example, most inborn errors of metabolism (i.e. phenylketonuria) qualify for inclusion into SMPDB. Likewise diseases such as diabetes,

metabolic syndrome, hyperlipidemia, various endocrine disorders as well as several kidney, liver and gastrointestinal disorders would also qualify. On the other hand, most cardiovascular diseases, pulmonary diseases, infectious diseases, cancers and many neuromuscular diseases would not. In a similar vein, drug pathways in SMPDB are defined as those pathways describing either the metabolism or mode of action of small-molecule drugs at a molecular, cellular or physiological level. Likewise, signaling pathways in SMPDB are defined as those pathways or physiological processes where signaling events are primarily mediated by small-molecule metabolites as opposed to proteins. This would include process such as neural or synaptic transmission, muscle activation, second-messenger signaling and small-molecule endocrine signaling. SMPDB signaling pathways would therefore not include most protein- or phosphoprotein-mediated signaling events.

An example of a typical SMPDB pathway (bile acid biosynthesis) is shown in Figure 1. As seen in this figure, all SMPDB pathways explicitly include the chemical structure of the major metabolites (or drugs) in each pathway. In addition, the cellular locations (membrane, cytoplasm, extracellular, mitochondrion, nucleus, peroxisome, etc.) of all metabolites and the enzymes or proteins involved in their processing are explicitly illustrated. Likewise, the quaternary structures (if known) and cofactors associated with each of the pathway proteins are also shown. If some of the metabolic processes occur primarily in one organ or in the intestinal microflora (as in bile acid biosynthesis), these features are also illustrated. We believe the inclusion of this 'structural' information (i.e. cellular, chemical and physiological) is one of the more unique and useful features of SMPDB. In particular, this kind of information provides both context and clarity to a number of metabolic processes that are highly dependent on certain key organs, organelles or chemical frameworks.

The SMPDB interface is largely modeled after the interface used for DrugBank (14) and the HMDB (15), with a navigation panel for browsing, searching and downloading the database. Below the navigation panel is a simple text query box that supports general text queries of the entire textual component of the database. Mousing over the Browse button allows users to choose between two browsing options, SMP-BROWSE and SMP-TOC. SMP-TOC is a scrollable hyperlinked table of contents that lists all pathways by name and category. SMP-BROWSE is a more comprehensive browsing tool that provides a tabular synopsis of SMPDB's content using thumbnail images of the pathway diagrams, textual descriptions of the pathways, as well as lists of the corresponding chemical components and enzyme/protein components. This browse view allows users to casually scroll through the database, select different pathway categories or re-sort its contents. Clicking on a given thumbnail image or the SMPDB pathway button brings up a full-screen image for the corresponding pathway. Once 'opened' the pathway image may be expanded by clicking on the Zoom button located at the top and bottom of the image or the magnifying-glass icons in the Highlight/Analyzer box on the right of the image. A pathway legend link is also available above the Zoom button. At the top of each image is pathway synopsis contained in a yellow box, while at the bottom of each image is a list of references. On the right of each pathway image is a gray–green Highlight/Analyzer tool with a list of the key metabolites/drugs and enzymes/proteins found in the pathway. Checking on selected items when in the SMP-Highlight mode will cause the corresponding metabolite or protein in the pathway image to be highlighted with a red box. Entering concentration or relative expression values (arbitrary units) beside compound or protein names, when in the SMP-Analyzer mode, will cause the corresponding metabolites or proteins to be highlighted with differing shades of green or red to illustrate increased or decreased concentrations.

As with most pathway databases, all of the chemical structures and proteins/enzymes illustrated in SMPDB's diagrams are hyperlinked to other online databases or tables. Specifically, all metabolites, drugs or proteins shown in the SMP-BROWSE tables or in a pathway diagram are linked to HMDB, DrugBank or UniProt (16), respectively. Therefore, clicking on chemical or protein image will open a new browser window with the corresponding DrugCard, MetaboCard or UniProt table being displayed. By hyperlinking to these particular databases, SMPDB is able to provide considerably more information about its small molecules and its proteins than any other pathway database. Indeed, SMPDB's UniProt links provide an average of 40 data fields about each protein, including nomenclature, function, reaction and structural information, while SMPDB's DrugBank/HMDB links provide an average of 100 data fields about each compound, including detailed descriptions, extensive nomenclature information, comprehensive physico-chemical data, reference NMR and MS spectra as well as extensive information about tissue or biofluid locations and concentrations.

SMPDB's Search menu offers users a choice of searching the database by chemical structure (ChemQuery), text (TextQuery), sequence (Sequence Search) or multiple identifiers (SMP-MAP). The ChemQuery option allows users to draw (using MarvinSketch applet) or write (using a SMILES string) a chemical compound and to search SMPDB for drugs and metabolites similar or identical to the query compound. The TextQuery button supports a more sophisticated text search (partial word matches, case sensitive, data field selection, Boolean text searches, misspellings, etc.) of the text portion of SMPDB. The Sequence Search button allows users to conduct BLASTP (protein) sequence searches of the protein sequences contained in SMPDB. Both single and multiple sequence BLAST queries are supported. The most powerful search option in SMPDB is SMP-MAP, which offers both multi-identifier searches as well as 'Omic' (transcriptomic, proteomic or metabolomic) mapping. Through a pull-down menu located at the top of the query textbox, SMP-MAP allows users to select the type of 'Omic' data (chemical names, HMDB IDs, DrugBank IDs, KEGG IDs, gene names, protein names, UniProt IDs, GenBank IDs, Agilent IDs or

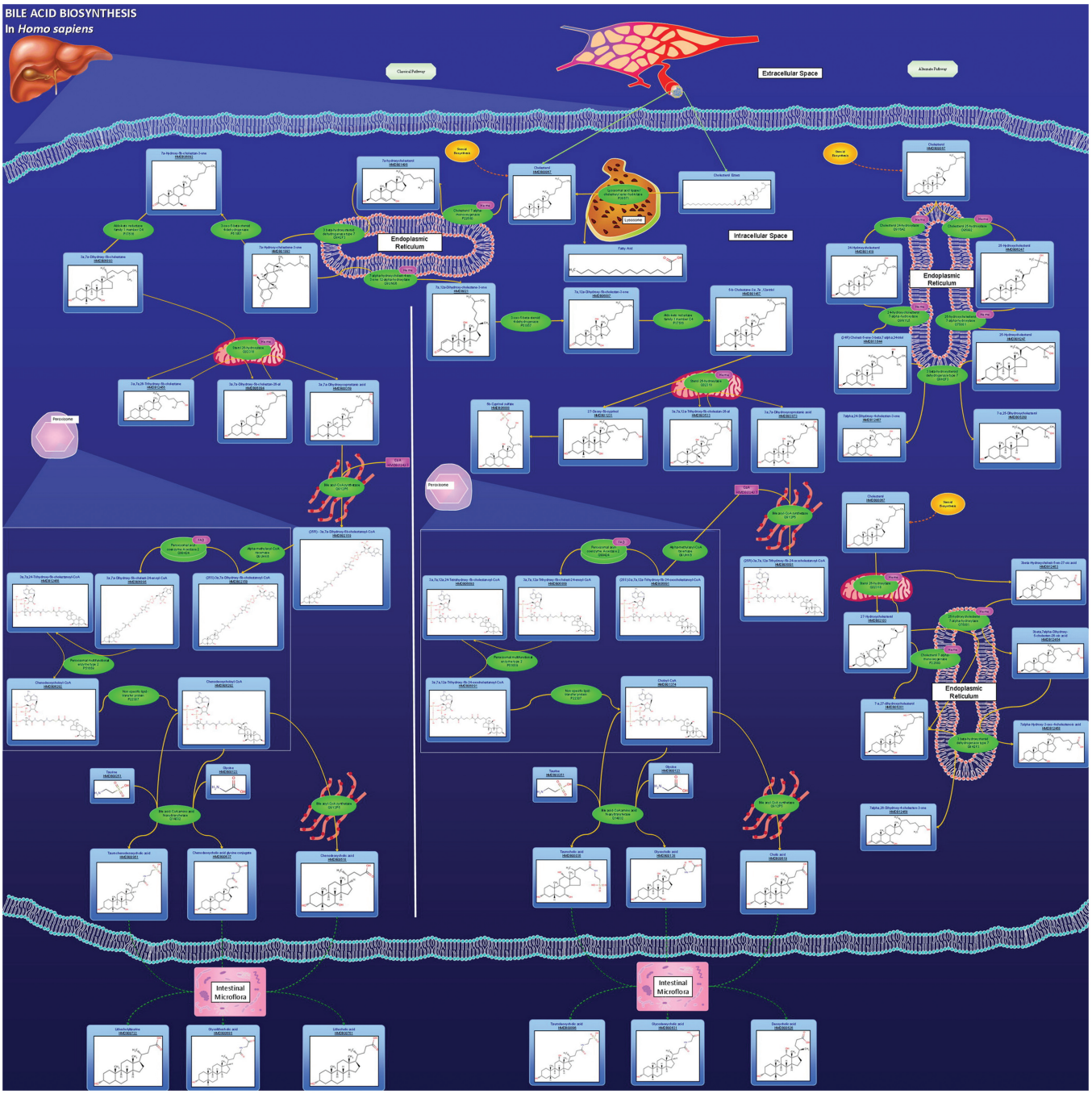

Figure 1. A screenshot of the Bile Acid Biosynthesis Pathway from SMPDB. This figure illustrates the typical design and layout of an SMPDB pathway diagram with the chemical structures, protein quaternary structures, cofactors, cellular locations, cellular structures and key organs all being explicitly shown.

Affymetrix IDs), then paste in a list of the identifiers and to have a table generated of appropriately highlighted pathways containing those components. The input lists may be pasted into the search box in a single column (name or identifiers only) or a double column (identifiers + concentration data) format. The resulting table, like the SMP-BROWSE table, displays a thumbnail image of the matching pathways along with the list of matching components (metabolites, drugs, proteins, etc.). The table is ordered by the number of matches and a significance score (calculated via a hypergeometric function), with the pathway having the most matches being at the top. Clicking on the thumbnail image or the SMPDB pathway button brings up a full-screen image for the corresponding pathway with all the matching components (metabolites, drugs, proteins, etc.) highlighted in red. Concentration data are displayed using a red-to-green gradient-color intensity scheme.

Highlighting is editable through the SMP-Highlight/Analyzer tool located beside the pathway image.

Through SMP-MAP, SMPDB is able to support pathway elucidation and pathway discovery in a variety of clinical 'omics' studies. For instance, users may assemble lists of significantly altered metabolites (or metabolite identifiers), genes (or gene identifiers) or proteins (or protein identifiers) from clinical metabolomic, microarray or proteomic studies. The resulting lists, scores and color-coded pathways can allow the identification or diagnostic confirmation of certain diseases, the association of certain 'omic' signatures with specific pathways or the discovery of pathway and/or disease associations that had not previously been detected. By allowing users to explore each high-scoring pathway individually, SMP-MAP may also allow new biochemical insights to be gained or suggest new confirmatory clinical tests to be performed.

In addition to the Browse and Search options, SMPDB also offers background data, citation information, references, links to other databases and statistical information under its 'About' menu. Users may also download sequence data, structure data and all of the image pathway data (as PowerPoint or JPG) images by navigating through the 'Download' menu.

QUALITY ASSURANCE AND CURATION

Each pathway diagram in the SMPDB is initially drawn using Microsoft PowerPoint using a palette of specially drawn icons to facilitate consistent illustration of organs, membranes, organelles, DNA, RNA, proteins, protein complexes, co-factors and chemical structures. A legend showing all of the standard SMPD icons and their meanings is given under the 'About' section in the SMPDB menu bar. Once drawn, the pathway image is converted to a JPG format and the chemical structures, proteins and co-factors are manually image mapped using the DreamWeaver software package. Three different-sized JPG images are prepared to facilitate rapid and consistent image zooming. All protein names, metabolite names, drug names, UniProt IDs, HMDB IDs and DrugBank IDs associated with each pathway are entered into a MySQL database to facilitate searching and hyperlinking. All pathways in the SMPDB have been drawn using a standard operating procedure (SOP) with a checklist of features that each of the pathway artists and pathway annotators were required to follow. This SOP and checklist is also provided in the 'About' menu. To ensure that the pathways have been knowledgably illustrated, all of the SMPDB pathway artists and annotators were required to have degrees in biology, biochemistry or bioinformatics. To check that the pathway illustrations (especially for metabolic and small molecule signaling pathways) were as correct and as comprehensive as possible, the SMPDB pathway artists and annotators were also required to consult a variety of sources including biochemistry textbooks, the Boerhinger–Manheim wall chart (2), OMIM (17), Wikipedia, KEGG (5), HumanCyc (7), Reactome (8), the Medical Biochemistry Page, UniProt (16), the HMDB (15), DrugBank (14) and PharmGKB (11). This allowed the annotation team to identify and consolidate pathway nomenclature, key pathway components, critical reactions, cellular compartments and key organs or organelles. Inconsistencies and errors were resolved by comparing data from multiple sources or finding review papers that clarified the situation. Pathway layouts were also 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 could be 'copied' from any other pathway diagram in any other database.

Because of the dearth of online drug and disease pathways, SMPDB's metabolic disease pathways and drug pathways were generated independently of any pathway database. Instead, relevant information was gathered from various medical and pharmacology textbooks, relevant encyclopedias, as well as several online databases such as OMIM (17) and DrugBank (14). This textual data were used to both design and render the pathway diagrams. The same SOPs for generating the metabolic and metabolite signaling pathways were used for generating SMPDB's disease and drug pathways. Reference information for every SMPDB pathway is provided in the pathway description textbox located below each pathway image. As an additional layer of quality assurance, all of the pathway diagrams in the SMPDB have been inspected and corrected by two or more curators having PhDs in biochemistry or physiology.

UNIQUE FEATURES AND COMPARISONS TO OTHER PATHWAY DATABASES

While there are several well-known metabolic pathway databases (KEGG, HumanCyc, Reactome, BioCarta, etc.), we have generally found that for the purposes of clinical 'omics' studies, the display formats, query options, information content and pathway coverage in these databases were often insufficient, sometimes incorrect or occasionally absent. In developing SMPDB, we not only tried to improve upon these shortcomings but also to build on some of the strengths of existing databases. We also endeavored to add content that is not normally found in other pathway databases. In particular, of the 364 pathways in SMPDB, 281 (or 72% of SMPDB's content) are unique. More specifically, 11/13 metabolite signaling pathways, 4/70 metabolic pathways, 154/168 drug pathways and 112/113 metabolic disease pathways depicted in SMPDB are not depicted in any form by KEGG, Reactome, EHMN, WikiPathways, HumanCyc, BioCarta, PharmGKB or any other database. Indeed, SMPDB is currently the only pathway database that includes significant numbers of metabolic disease pathways (>110) and drug pathways (>160). In addition to providing a large number of novel pathways, SMPDB also adds a significant amount of new information content, including more than 30 000 words of original text describing each pathway in the context of human physiology and human biochemistry. Furthermore,

each drug or metabolite in SMPDB is hyperlinked to detailed descriptions of that molecule, including extensive nomenclature information, comprehensive physico-chemical data, thousands of reference nuclear magnetic resonance (NMR) and mass spectrometry (MS) spectra as well as extensive information about tissue or biofluid locations and concentrations (~100 data fields per compound). In addition to this textual content, SMPDB also offers a significant amount of new and useful graphical content, including the depiction of the relevant organs, cellular locations, organelles, membrane boundaries, protein quaternary structures, cofactors and other cellular features in its pathway diagrams.

With regard to its interface design, SMPDB also offers a number of new or unique features. In particular, SMPDB uses thumbnail images to facilitate pathway viewing and browsing, it uses a scrollable table to display pathways and pathway synopses, and it employs a unique checkbox Highlight/Analyzer tool to allow users to interactively highlight and color multiple metabolites, drugs and/or proteins on its pathway images. SMPDB also uses a graphical structure-searching applet to enable sophisticated drug or metabolite similarity searches. As with some of the more fully developed pathway databases, SMPDB also provides protein/metabolite lists for each pathway, it supports advanced text and sequence queries and it allows pathway mapping from protein, gene and metabolite lists. While space does not permit a detailed comparison against all existing pathway databases, it is perhaps instructive to compare SMPDB to six of the larger and more established resources: KEGG, Reactome HumanCyc, BioCarta, EHMN and WikiPathways/GenMAPP.

This comparison is summarized in Table 1 where we have used a number of general features or criteria to make our assessment. Some of these criteria definitions may need further elaboration. In particular: 'Information provided on pathway entities' is defined as providing hyperlinks to pages that give additional detail on protein/drug/metabolite sequences, functions, properties, reactions, structure, concentrations or spectra. 'Supports advanced text search' is defined as supporting field-specific searches, wild-card queries, Boolean searches, synonym searches, mis-spellings, text sorting or other kinds of complex textual queries beyond simple text matching. 'Component lists available' is defined as providing easily accessed, plain text or hyperlinked textual lists of all the genes, proteins, drugs and/or metabolites displayed in the pathway. In addition to the information in Table 1, we have also elaborated on the comparisons for four of the databases (KEGG, HumanCyc, Reactome and BioCarta) in the following paragraphs.

KEGG, with 330 reference pathways, is considered to be the 'gold standard' for most pathway databases because of its comprehensiveness and its breadth of organism-specific coverage. Of KEGG's 158 metabolic pathways (Table 1), 73 are relevant to humans or other mammals. Interestingly, several key metabolic pathways in mammals are actually missing from KEGG (i.e. the malate-aspartate shuttle, and electron transfer). As the pathway diagrams in KEGG are designed to be very 'generic' they display no organ data, no cellular structure information, no protein superstructure data, no chemical structure information (except through hyperlinks) and no gene or protein names (only EC numbers). While KEGG does have 35 disease pathways, most are for cancer, neurological or immune diseases and only three of these relate to small-molecule metabolites or metabolic diseases. KEGG's drug pathways are limited to showing only drug development or drug similarity as opposed to drug action

Table 1. Comparison of SMPDB to KEGG, HumanCyc, Reactome, BioCarta, EHMN and WikiPathways/GenMAPP

Feature	SMPDB	KEGG	Reactome	HumanCyc	BioCarta	EHMN	WikiPathways
Number of metabolic pathways	70	73 (for humans)	64 (for humans)	317 (total), 238 (conf)	55	70	44 (for humans)
Number of disease pathways	113	35	3	0	24	0	4
Number of drug action pathways	168	0	3	3	10	0	4
Provides multiple organism pathways	No	Yes	Yes	No	Yes	No	Yes
Chemical structures shown in diagrams	Yes	No	Some	Yes (when zoomed)	Some	No	No
Protein 4° structures shown in diagrams	Yes	No	No	No	Some	No	No
Cell structures shown in pathway diagrams	Yes	No	No	No	Yes	No	No
Organs shown in pathway diagrams	Some	No	No	No	No	No	No
Descriptions of pathways provided	Yes	No	Yes	Yes	Yes	No	Some
Pathway images are hyperlinked	Yes	Yes	Limited	Yes	Yes	No	No
Pathway images are zoomable	Yes	No	No	Yes	No	Yes	Limited
Information provided on pathway entities	Detailed	Modest	Limited	Moderate	Limited	Limited	Limited
Supports simple text search	Yes	Yes	Yes	Yes	No	Yes	Yes
Supports advanced text search	Yes	No	Yes	Yes	No	No	No
Supports sequence searching	Yes	No	No	Yes	No	No	No
Supports graphical chemical structure search	Yes	No	No	No	No	No	No
Supports chemical expression mapping	Yes	No	Yes	Yes	No	No	No
Supports gene/protein expression mapping	Yes	No	Yes	Yes	No	No	No
Downloadable	Yes	Yes	Yes	Yes	No	No	Yes
Component lists available	Yes	No	Yes	No	Proteins only	Yes	Yes
BioPax, CellML or SBML compatible	No	Partial	Yes	Yes	No	Yes	No

or drug mechanism. While KEGG does provide very useful annotations (via hyperlinks) for the compounds shown in its pathways, it does not provide descriptive summaries of these pathways nor does it support the visual display of gene/protein/metabolite concentrations. As yet, KEGG does not provide protein/metabolite lists for each pathway nor does it support graphical structure queries or chemical structure similarity searches.

Similar to KEGG, the Reactome database provides 1000s of metabolic and signaling pathway data sets for many model organisms. Of these, 64 pathways are associated with human metabolism, three with human diseases and three with drug action or drug metabolism. Through its Reaction Map interface, Reactome is able to provide users with low-resolution pathway maps (similar to the thumbnail images used by the SMPDB browser) that allow users to interactively navigate through the database. Like SMPDB, Reactome provides extensive pathway or reaction descriptions along with hyperlinks to several external databases. Unlike SMPDB, Reactome does not display organ data, cellular compartment data, cellular organelle information, protein complex information or protein/gene names in its pathway diagrams. Likewise, Reactome has very few disease or drug pathways. On the other hand, Reactome, like SMPDB, has a 'Skypainter' feature allows users to paste in a list of genes or gene identifiers and to 'paint' the Reactome reaction map in a variety of ways. Unlike SMPDB, Reactome does not support chemical structure or sequence queries.

The HumanCyc database contains 349 pathways, including 29 superpathways (supersets of many of the other 320 pathways in the database). Two hundred and thirty-eight of these pathways are confirmed, meaning they have 'evidence glyphs' indicating 50% or more of the reactions have some evidence of occurring in humans. All pathways in HumanCyc can be 'zoomed-in' to display chemical structures, EC numbers and protein names, similar to SMPDB. Similar to SMPDB, HumanCyc supports advanced text searches as well as sequence searches. HumanCyc's metabolic pathways are well referenced and generally well described, although most descriptions are given in the context of bacterial or plant metabolism. The images in HumanCyc do not display organ data, cellular compartment data, cellular organelle information or protein complex information. Currently, HumanCyc does not have any disease pathways and it provides only three drug pathways. While HumanCyc does support the visual display of gene/protein/metabolite concentrations using its own 'OmicsViewer', the display is small and sometimes difficult to interpret—especially for metabolomics applications. Unlike SMPDB, HumanCyc does not provide protein/metabolite lists for each pathway nor does it support chemical structure similarity searches.

The BioCarta database, with 360 pathways, is probably the most visually sophisticated database of the four pathway databases described here. Like SMPDB, not only most of its pathways depict cell, protein and chemical structure information, but also all of its pathways are well annotated with very detailed descriptions. BioCarta also contains a large number (>250) of protein signaling pathways and many other macromolecular interaction processes. However, BioCarta only has a modest number (55) of pathways that are devoted to small molecule metabolism and an even smaller number (<35) of pathways that are devoted to disease or drug action pathways. Very few of these disease or drug pathways overlap with those in SMPDB. As BioCarta is a community-annotated/generated database, its collection of pathways is somewhat haphazard and largely dependent on community interest. Likewise, BioCarta's querying and display tools are very limited compared to most other pathway databases.

LIMITATIONS AND FUTURE PROSPECTS

Compared to most other pathway databases, SMPDB is really a niche database that is specific to just one (albeit important) organism—namely humans. In this regard, SMPDB certainly lacks the breadth of metabolic coverage offered by KEGG, Reactome or the 'Cyc' family of databases. Likewise, because SMPDB is focused on small molecules, it does not include the key protein signaling pathway information found in other databases, such as KEGG, BioCarta or TransPath. As a consequence, SMPDB is not—nor will it ever be—very useful for those interested in comparative metabolic studies, protein network analysis, metabolic engineering, protein signal transduction or metabolic evolution. Because of its narrow focus and restricted mandate, SMPDB should not be thought of as a replacement for existing pathway databases. Rather, the information in SMPDB should be viewed as complementary to what is already out there.

While SMPDB is essentially a database of small-molecule pathways, it is also a database of scientific artwork. The art, in this case, corresponds to our best understanding and our best efforts at representing the complex chemical and biochemical processes found in humans. As such, SMPDB is not without flaws. Certainly, with any illustration or piece or art, there are other (potentially better) ways of rendering the process or in capturing its essence. Likewise, because art is very much a personal choice, it is likely that the style, layout and color scheme adopted by SMPD may not agree with every user. For instance, the display of all chemical structures (which vary widely in size) may make some SMPDB pathway images a little difficult to interpret. However, because SMPDB is primarily designed to serve the metabolomics community, this made chemical structure rendering an essential part of its construction. Other aspects, such as the use of a dark blue background (to represent aqueous conditions) and thin directional arrows (to reduce clutter), may not appeal to all users. However, the fact that all of SMPDB's PowerPoint images are downloadable, certainly should allow users to re-render or re-color the pathways to their liking.

While considerable effort and many revisions went into developing both the current pathways and standard pathway drawing protocols, SMPDB is far from uniform. It is worth noting that SMPDB was assembled from

multiple artists, and this means there will always be some inconsistency in how data are schematically represented and how much detail is displayed in any given figure. Again, these issues can be addressed if users provide feedback to SMPDB's annotation team or if they choose to download and edit SMPDB images on their own.

Another important limitation to SMPDB is the fact that its pathways are not (yet) represented in BioPax or SBML format (18). Efforts will be made in the future to convert most SMPDB images to SBML and SimCell (19) compatible formats, but it is important to note that SMPDB was designed to be more of a human-readable database rather than a machine-readable resource. Consequently, most of our computational effort has gone into to providing powerful searching, querying and viewing tools to facilitate user interactivity.

Like many existing pathway databases, SMPDB is still a work in progress. We are currently adding two to three new pathways a week and are constantly revising or improving the renderings of existing pathways images based on new data or on user feedback. While the number of metabolic pathways is unlikely to change substantially, it is expected that the total number of disease and metabolic signaling pathways will eventually exceed 200, and the number of drug action pathways will certainly exceed 400. In addition, to the planned expansion in size, SMPDB's metabolic and disease pathways will be formally linked to the next release of the HMDB, while its drug action pathways will be linked the next release of DrugBank. SMPDB is also being integrated into a virtual reality 'CAVE' environment for interactively visualizing and modeling human metabolism (20). Overall, we believe SMPDB is a useful addition to the current collection of pathway databases and we are hopeful that some of its more innovative or useful ideas will eventually find their way into other online pathway databases.

FUNDING

The Alberta Ingenuity Fund (AIF); Alberta Advanced Education and Technology (AAET); the Canadian Institutes for Health Research (CIHR); Genome Alberta, a division of Genome Canada. Funding for open access charge: Canadian Institutes for Health Research.

Conflict of interest statement. None declared.

REFERENCES

1. Michal,G. (1998) On representation of metabolic pathways. *Biosystems*, **47**, 1–7.
2. Michal,G. (1968) *Biochemical Pathways Wall Chart*. Boehringer Mannheim GmbH, Germany.
3. Ideker,T., Galitski,T. and Hood,L. (2001) A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.*, **2**, 343–372.
4. Suderman,M. and Hallett,M. (2007) Tools for visually exploring biological networks. *Bioinformatics*, **23**, 2651–2659.
5. Okuda,S., Yamada,T., Hamajima,M., Itoh,M., Katayama,T., Bork,P., Goto,S. and Kanehisa,M. (2008) KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res.*, **36**, W423–W426.
6. Karp,P.D., Riley,M., Paley,S.M. and Pelligrini-Toole,A. (1996) EcoCyc: an encyclopedia of Escherichia coli genes and metabolism. *Nucleic Acids Res.*, **24**, 32–39.
7. Caspi,R., Foerster,H., Fulcher,C.A., Kaipa,P., Krummenacker,M., Latendresse,M., Paley,S., Rhee,S.Y., Shearer,A.G., Tissier,C. *et al.* (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **36**, D623–D631.
8. 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. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33**, D428–D432.
9. Pico,A.R., Kelder,T., van Iersel,M.P., Hanspers,K., Conklin,B.R. and Evelo,C. (2008) WikiPathways: pathway editing for the people. *PLoS Biol.*, **6**, e184.
10. Ma,H., Sorokin,A., Mazein,A., Selkov,A., Selkov,E., Demin,O. and Goryanin,I. (2007) The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol. Syst. Biol.*, **3**, 135.
11. Sangkuhl,K., Berlin,D.S., Altman,R.B. and Klein,T.E. (2008) PharmGKB: understanding the effects of individual genetic variants. *Drug Metab. Rev.*, **40**, 539–551.
12. Wishart,D.S., Lewis,M.J., Morrissey,J.A., Flegel,M.D., Jeroncic,K., Xiong,Y., Cheng,D., Eisner,R., Gautam,B., Tzur,D. *et al.* (2008) The human cerebrospinal fluid metabolome. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, **871**, 164–173.
13. Wishart,D.S. (2008) Applications of metabolomics in drug discovery and development. *Drugs R D*, **9**, 307–322.
14. Wishart,D.S., Knox,C., Guo,A.C., Shrivastava,S., Hassanali,M., Stothard,P., Chang,Z. and Woolsey,J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–D672.
15. Wishart,D.S., Tzur,D., Knox,C., Eisner,R., Guo,A.C., Young,N., Cheng,D., Jewell,K., Arndt,D., Sawhney,S. *et al.* (2007) HMDB: the Human Metabolome Database. *Nucleic Acids Res.*, **35**, D521–D526.
16. UniProt Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.
17. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
18. Strömbäck,L. and Lambrix,P. (2005) Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics*, **21**, 4401–4407.
19. Wishart,D.S., Yang,R., Arndt,D., Tang,P. and Cruz,J. (2005) Dynamic cellular automata: an alternative approach to cellular simulation. *In Silico Biol.*, **5**, 139–161.
20. Soh,J., Turinsky,A.L., Trinh,Q.M., Chang,J., Sabhaney,A., Dong,X., Gordon,P.M., Janzen,R.P., Hau,D., Xia,J. *et al.* (2009) Spatiotemporal integration of molecular and anatomical data in virtual reality using semantic mapping. *Int. J. Nanomed.*, **4**, 79–89.