Published in final edited form as:

Comb Chem High Throughput Screen. 2013 March ; 16(3): 189–198.

AN OVERVIEW OF COMPUTATIONAL LIFE SCIENCE DATABASES & EXCHANGE FORMATS OF RELEVANCE TO CHEMICAL BIOLOGY RESEARCH

Aaron Smalter Hall, Yunfeng Shan, Gerald Lushington, and Mahesh Visvanathan*

Bioinformatics Core Facility, University of Kansas, Lawrence, Kansas, 66047, United States

Abstract

Databases and exchange formats describing biological entities such as chemicals and proteins, along with their relationships, are a critical component of research in life sciences disciplines, including chemical biology wherein small information about small molecule properties converges with cellular and molecular biology. Databases for storing biological entities are growing not only in size, but also in type, with many similarities between them and often subtle differences. The data formats available to describe and exchange these entities are numerous as well. In general, each format is optimized for a particular purpose or database, and hence some understanding of these formats is required when choosing one for research purposes. This paper reviews a selection of different databases and data formats with the goal of summarizing their purposes, features, and limitations. Databases are reviewed under the categories of 1) protein interactions, 2) metabolic pathways, 3) chemical interactions, and 4) drug discovery. Representation formats will be discussed according to those describing chemical structures, and those describing genomic/proteomic entities.

Keywords

Databases; Data Exchange Formats; Proteomics; Drug Discovery

1. INTRODUCTION

Computational Life Science (CLS) is an interactive approach that facilitates the study of biological systems by integrating molecular profiling data (genomics, transcriptomics, proteomics and metabolomics) with genetic information to get a deeper understanding of their functional relationships at the system level. In this endeavor, separation science plays a key role in collecting sufficient information on biological processes of interest at various levels as shown in figure 1. Modern data management tools are also very important to analyze, compare and evaluate large data sets. The ability to combine a comprehensive bioanalytical platform with systems biology approaches has great potential to decipher complex biological systems, including the capability to systematically follow the dynamics of biological processes from gene expression to phenotype analysis. With the assistance of

*To whom correspondence should be addressed: mvisvanathan@ku.edu; 785-864-3337.

state of the art separation and systematical analysis tools, CLS is expected to play a fundamental role in pharmaceutical and biotechnology research to serve current health care needs from drug discovery to personalized medicine [1].

In recent years, research in biology has generated a huge amount of data. There are numerous biological databases that maintain knowledge on genes, proteins, metabolites and other substances. There are also various commonly agreed formats for data presentation and data exchange. However, due to the complexity of such data it is challenging to standardize the representation and integration of larger data sets. The lack of full standardization hinders effective integration of results from cross-disciplinary collaborative studies.

Systematic approaches are needed to maintain and analyze biological data. In this post-genomics era the occurrence of new genomic and proteomic data models has also led to developing new systematic methods to study and analyze them. In CLS, synergistic applications of experiment, theory and modeling towards understanding biological processes as a whole system require an integrated software environment. The software environment should include a comprehensive variety of capabilities like access to various databases, tools for formalized description of biological systems, tools for visualization and performing simulations. Managing and understanding the complexity of modeling for genomic and proteomics experiments is a challenge, towards which numerous tools are being developed to facilitate understanding of biological pathways relevant to drug discovery and development.

Naturally, there is some redundancy and overlap between these databases. Biological entities are highly interconnected, with chemicals, proteins, DNA, and RNA all being inter-related through a variety of interactions, reactions and pathways. Figure 2 provides a simplified depiction of this situation. Because of this integrated nature, databases describing these entities often cover more than a single entity/relationship and hence may serve more than one purpose. Some broad initiatives such as NCBI's Entrez search engine, or the Kyoto Encyclopedia of Genes and Genomes (KEGG) seek to unify databases cataloging most (or even all) of the different biological entities and their relationships. Despite establishment of some monolithic resources, specialized databases are still being created to address important research issues. Given these specialized applications, these databases do not abide by any community standard for database schema structure and datatype, format or minimum information content. Nonetheless, there would be great value in combining different data sources and performing biological annotations corresponding to pathway models [2].

This paper is devoted to reviewing a diverse selection of these databases and exchange formats so that researchers will be better informed in selecting specific resources for their work. The databases examined belong to four broad categories: protein-protein interactions, biological pathways, chemical interactions, and drug discovery. The data exchange formats will also be reviewed according to their specific groups. The table below details the database categories under review and provides some information on the entities described by them and some examples.

2. DATA RESOURCES IN CLS

Examining both the literature and the Internet results in a large and varied list of databases that contain interaction information for proteins, DNA, RNA and small organic molecules. The large number of such projects reflects the importance of this data to life sciences research. Many databases are packed with information, but the information is structured in such a way that it cannot be unambiguously matched to other databases. For example, some databases do not contain key data descriptors like sequence accession numbers, chemical compound numbers and PubMed [3] identifiers for publication references. This diminishes the usefulness of the information, since it is difficult to tie it to other knowledge, which is required on a large scale for achieving broad contextual understanding. It is critical that these projects move towards sound database principles when describing data such that it may be computed upon unambiguously and precisely.

2.1. PROTEIN INTERACTION DATABASES

During the last 20 years there has been an increasing interest in applying databases to biological studies. Markowitz et al [4] discussed a few criteria for evaluation and comparison of molecular biology databases. A substantial amount of research has focused on genetic codes, amino acid sequences of proteins and 3-D protein structures that show usage of database in biological research [5–7]. While protein interaction databases have been summarized in detail elsewhere (e.g., Fuentes et al. [8]) we present the various available biological and mathematical data sources that are currently available.

BIND—(Biomolecular Interaction Network Database) [9] stores descriptions of interactions, molecular complexes and pathways.

MIPS—(Munich Information center for Protein Sequences; <http://mips.gsf.de>) [10] provides whole genome protein sequence-based information for various model organisms, integrating a number of databases (each devoted to a specific organism or contextual focus) including:

- CYGD – yeast genome, discussed in more detail below
- MNCDB – Neurospora crassa genome
- NGFN – German Human Genome Project
- MPPI – Mammalian Protein-Protein Interactions
- SIMAP – FASTA homologies
- MATDB – Arabidopsis thaliana
- MOsDB – rice genome
- SPUTNIK – plant ESTs
- PEDANT – comprehensive set of genomes

While distinct, these database are all comprehensive organism genome resources. MIPS contains both automatic and manually curated records, with systematic classification schemes and functional protein annotations.

DIP—(Database of Interacting Proteins) [11] stores protein-protein interactions, including physical associations and chemical reactions and chemical states of those proteins. DIP represents interactions via a binary interaction scheme, depicted schematically with a graph abstraction and a visual navigation tool. DIP does enforce a formal grammar for data specification, but does allow description of the interacting proteins, experimental methods underlying the interaction determination, quantifies the dissociation constant for physical associations, reports the amino acid residue ranges for the interaction site and provides references for the interaction.

CYGD—(Comprehensive Yeast Genome Database; <http://mips.gsf.de/genre/proj/yeast/>) [12] summarizes current knowledge (i.e., chromosomes/genes, and functional interaction therein) regarding the 6,200+ open reading frames (ORFs) within the *Saccharomyces cerevisiae* genome, with each ORF having a length of more than 99 amino acid residues. CYGD is a component of MIPS; it can be searched or browsed by chromosome using a web-based interface, and raw data can be downloaded via FTP.

EcoCyc—(<http://ecocyc.org/>) is an NIH-funded database reporting metabolic and signaling pathway information for *Escherichia coli* [13]. EcoCyc employs an object-oriented data structure containing literature-curated gene associations, such as transcriptional and regulatory information, pathways, reaction participation, and Gene Ontology annotations. All associations are backed by literature references, and EcoCyc provides a number of metabolic, transcriptional, and regulatory diagrams for visualization.

MINT—(Molecular Interaction Network Database) [14] is a molecular interactions database assembled from the literature and manually input. In addition to a simple relational schema for representing binary relations, MINT records information about protein post-translational modifications, experimental metadata, cellular location, pathway participation and known complexes. MINT to supports experimental verification of protein-protein interactions, and provides detailed interaction data including kinetic and binding constants and associative domain annotation. MINT uses an automated software system to scan abstracts and suggest literature studies to be manually curated by domain experts. As of March 2011, MINT reported 90290 interactions among 31870 proteins.

2.2. PATHWAY DATABASES

These databases organize information about metabolic reactions and the distinct pathways formed by the interplay between those reactions. These databases cover pathways for an increasing variety of organisms. Some of the most widely used pathway databases are listed below.

REACTOME—(www.reactome.org) [15] is an open-source, peer-reviewed pathway database whose pathway annotations are curated by domain experts and linked to many other life science resources, including the NCBI Entrez Gene, Ensembl, UniProt, KEGG Compound and ChEBI small molecule databases, the UCSC and HapMap Genome Browsers, PubMed, and Gene Ontology. The REACTOME data model centers on reactions among entities (nucleic acids, proteins, complexes and small molecules), which establishes

an interaction network that can be subdivided into pathways for humans and 20 non-human species, while collaborative extensions have been pursued to address drosophila, Arabidopsis and other model systems. REACTOME provides analysis and comparison tools that permit predicted pathways in one organism to be mapped to pathways in another species.

KEGG—(Kyoto Encyclopedia of Genes and Genomes) [16] aims to comprehensively describe and relate biological entities by compilation of several interrelated databases addressing different aspects of biology. One KEGG component is the KEGG Pathway database, which stores and describes metabolic reactions that can be retrieved as manually curated graphical diagrams. These diagrams are highly interconnected and link all known metabolic reactions. Each pathway reactants in the database is linked to detailed descriptive records.

BioCyc—(<http://biocyc.org>) [17] is a suite of 1129 databases (ca. March 2011), each of which reports genome and pathway characterization of a distinct organisms. The BioCyc databases are arranged hierarchically by the level of curation and data confidence. The top tier of databases is the most intensively curated with literature references and commentary, and includes records for some important, heavily studied model organisms such as *Escherichia coli*, *Arabidopsis thaliana*, and *Saccharomyces cerevisiae*. The next tier focuses on models that are largely computationally derived, but have received some curation and validation from domain experts. The last tier contains purely computationally derived records with no manual curation. Currently there are four Tier 1 databases, 32 tier 2 databases, and 1084 tier 3 databases. Access to these resources is facilitated by centralized query and visualization tools for omics data analysis, comparative genomics and pathway analysis. Links are also provided to other biological databases containing protein and nucleic-acid sequence data, bibliographic data, protein structures, and organismic descriptions.

2.3. MATHEMATICAL MODELING DATABASES

Since interaction knowledge is constantly evolving, recent efforts have been directed toward developing and maintaining quantitative mathematical pathway models based on minimal direct biological information. This new dimension of molecular cell biology entails assembly of quantitative associations based on mining primary literature studies and curated interaction databases. However, such mathematical pathway models cannot be searched based on specific biological knowledge, thus effective querying is best accomplished if both qualitative and quantitative information is combined and stored in a database. A number of databases have arisen to address this requirement, of which some recent examples are listed below.

Biomodels—(Biomodels.net) [18] reports published, peer-reviewed, quantitative models of biochemical and cellular systems, with manually curated annotations and cross-references to other relevant data resources. The resource enables precise specification of model components, and efficient model retrieval and visualization. A key Biomodels objective is to

ensure that all the models published in the public domain are made freely available for everyone.

DOQCS—(Database of Quantitative Cellular Signaling) [19] is a database documenting chemical kinetic models of signaling networks, described more in detail elsewhere [20]. The objective underlying the DOQCS project was to aggregate experimental the data and to facilitate collaborations between biologists and modelers to unravel key mechanistic aspects of cell signalling.

UniPathway—(<http://www.grenoble.prabi.fr/obiwarehouse/unipathway>) [21] is a database component of the UniProtKB/Swiss-Prot collaboration designed to store metabolic pathway records. It provides a controlled vocabulary for describing the pathways and their associated proteins. Links between protein records and metabolic reactions are manually curated by domain experts. UniPathway is also affiliated with the Open Biological Warehouse and those cross-references a large array of genomic, metabolic, structure/sequence, and ontology databases.

2.4. CHEMICAL INTERACTION DATABASES

With the increasing amount of high-throughput chemical biology screening results generated by centers and labs around the world, a number of databases have arisen to collect, categorize, and manage these results. Typically these databases consist of two components, 1) a large library of chemical structures, and 2) a set of assay results linking selections of chemicals compounds to observed bioactivities.

PubChem—(<http://pubchem.ncbi.nlm.nih.gov>) [22] is a large resource of chemical structure and bioactivity information organized into three affiliated databases: Compounds, Substances, and BioAssays. PubChem is maintained by the National Center for Biotechnology Information (NCBI), part of the NIH Molecular Libraries Program, and has become a ubiquitous resource for research in cheminformatics and related fields. As of January 2011, the Substances section of PubChem contains over 75 million records describing chemical mixtures, extracts, and complexes, while the Compounds section contains over 31 million records of well-characterized chemicals.

The compound and substance records describe not only the two-dimensional structures of chemicals, but also compound metadata, chemical properties and links to bioactivity results and other related compounds. These structures and other information can be obtained from PubChem in a variety of common formats such as XML or SDF. The BioAssays section of PubChem contains nearly 500,000 reports from high-throughput screening experiments, cataloging millions of bioactivity endpoints such as toxicity or target inhibition. BioAssay records contain descriptions of experimental protocols and other relevant information along with the specific activity results.

PubChem databases are publicly accessible through a web-based interface that supports a fairly broad range of query and analysis options, and the raw data and structures can be downloaded via FTP. PubChem allows voluntary deposition of new records by researchers, and while depositions are screened by curators, PubChem data is not exhaustively validated.

ChemBank—(<http://chembank.broad.harvard.edu/>) [23] is designed as not only a database, but also an environment for the analysis of small molecules and their bioactivity. It was created by the Chemical Biology Program of the Broad Institute at MIT and Harvard, with funding from the United States National Cancer Institute. ChemBank is an intensively curated resource with a focus on high-throughput screening data. As of August 2007 it contained information on more than 1.2 million unique chemical structures with from at least 2500 assays. In addition to structure and activity data, ChemBank also calculates more than 300 molecular descriptors and organizes bioassays hierarchically using metadata. It also offers a suite of analysis tools such as searching by similarity, descriptors, or substructures, as well as visualization of screening results and chemical-genetic profiles.

ChEBI—(Chemical Entities of Biological Interest; <http://www.ebi.ac.uk/chebi/>) [24], is a database of “molecular entities” which can refer to chemical structures or mixtures containing any number of components. Whole molecules as well as molecule fragments, complexes, or even individual atoms are indexed. The criterion for inclusion in the database is that all of the entities are natural products of organism biology, or can interfere in biological processes (such as toxins or synthetic drugs). Biomolecules that result from genetic processes (nucleic acids, proteins, etc.) are excluded from this database. ChEBI uses standard representations and formats from the International Union of Pure and Applied Chemistry (IUPAC) and Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB).

ChemDB—(<http://cdb.ics.uci.edu>) [25] is a publicly available database of small molecules. The distinguishing feature of this database is that is compiled from the compound collections maintained by over one hundred industry and public laboratories. Additionally, ChemDB includes computationally derived predictions and annotations such as solubility and three-dimensional structure. Currently the ChemDB database contains more than 4.1 million commercially available chemical compounds, with over 8.2 million isomers. Like other drug discovery databases, an important component of ChemDB is the analysis tools available to visualize and search chemicals and chemical reactions, among other other uses.

STITCH—(‘search tool for interactions of chemicals’, <http://stitch.embl.de/>) [26] is a database and also an information management system through which to analyze chemical interactions and metabolic pathways. STITCH draws upon reported crystal structure information, as well as experimentally obtained drug-target interactions. It can also generate predictions about chemical activity through comparisons using text mining and structural similarity. Again, STITCH provides a suite of tools to search and visualize chemical/protein interaction networks. The STITCH database contains records for over 68,000 chemical compounds, including 2,200 known drugs and their relationships with over 1.5 million genes in 373 genomes.

2.5. DRUG DISCOVERY DATABASES

While many of the previously discussed databases can certainly be used in drug discovery research, there exist several databases specifically designed with drug discovery in mind. The chemicals stored in these databases are often filtered to contain molecules that are

already FDA approved or drug-like to some extent. Additionally, these databases provide detailed information regarding drug targets as well as computational tools to guide the drug development process.

BioDrugScreen—(www.biodrugscreen.org) [27] is a combined web-based drug discovery database and analysis server. The database portion of the service consists of the Docked Proteome Interaction Network (DOPIN). The distinguishing feature of DOPIN is that in addition to storing chemical-protein and protein-protein interactions, it reports results from pre-computed docking simulations. These docking scores can be combined with customized scoring functions in order to rank chemical compounds and their possible targets. The scoring function can be arbitrarily defined to take into account structural information or other descriptors extracted from existing databases. BioDrugScreen also offers tools to evaluate the scoring functions and resulting predictions, and can also perform on demand docking of user-uploaded molecules against preprocessed targets from the PDB database [28].

DrugBank—(www.drugbank.ca) [29] is drug discovery database combining detailed target information with pharmaceutically relevant drug information. With regards to drug targets, DrugBank stores raw sequence data, as well as conformational structures and links to associated pathways and splice variants. The drug information records contain a number of properties relevant to drug discovery such as mode of action and pharmacokinetic profile. As of March 2011, Drugbank stores over 6800 records for drugs, with about 1400 of those drugs considered FDA-approved and about 5200 considered experimental. It also stores over 4400 protein sequences with links to interacting drugs. DrugBank can be browsed or searched using textual or similarity based queries. Records are publicly available for download. DrugBank was developed at the University of Alberta.

TTD—(The Therapeutic Target Database; <http://xin.cz3.nus.edu.sg/group/CJTDD/>) [30] is a drug-target database focused on therapeutic applications. It indexes protein and nucleic acid targets, annotating them with information about pathway participation, disease association, corresponding drugs and other relevant properties. These records are cross-referenced with other databases noting sequences, structures, binding affinities, and clinical results, among other details. TTD is created by the Bioinformatics and Drug Design Group at the National University of Singapore and currently contains over 1,900 known, clinical, and research targets and over 5,000 approved, clinical, and experimental drugs. This set of targets and drugs covers 61 classes of proteins and 140 classes of drugs. TTD allows searching via textual queries on multiple fields.

KiBank—(<http://kibank.iis.u-tokyo.ac.jp>) [31] is a web-based database for computer-aided drug design implemented in the University of Tokyo. It provides information on binding affinities, chemical and target protein structures. 2D and 3D structure visualizations are accessible for most compounds. KiBank is unique in that the structural representations stored are generated using a molecular simulation that effects energy minimization of the complex. This feature allows structure records from KiBank to be used directly for computational binding studies that rely on accurate three-dimensional structures. As the

name suggests, KiBank also stores Ki inhibition constants in order to evaluate protein-chemical interaction strengths. As of 2004, there were over 6000 records of binding affinity available in KiBank, manually curated from literature sources. Those binding affinities link a total of 142 protein targets to over 5,000 chemicals.

3. VARIOUS DATA EXCHANGE FORMATS IN CLS

To complement the diverse range of pathway and protein/chemical interaction databases and their varying data forms and specifications identified in the previous section, various different forms of mathematical models are available for pathway model characterization.

3.1 DATA FORMATS FOR GENOMICS AND PROTEOMICS

As the number of databases storing proteomic and genetic records increase, as well as the size of those databases, it is increasingly important to define standardized data exchange formats to facilitate collaboration between researchers. This subsection will review several data formats in common proteomics and genomics databases.

PSI MI—(The proteomics Standard Initiative Molecular Interaction) constitutes an attempt to develop community standards for data representation in proteomics and to facilitate comparison, exchange and verification. Universal adherence to these standards has not been achieved; however some important resources such as DIP and TransPath do implement PSI MI specifications. Currently in PSI MI format the data are structured around entry. An entry describes one or more interactions that can be considered as one unit. The entry contains following parts: Source, Availabilitylist, Experimentalist, Interactionlist, Interactorlist and Attributelist. The aim is to extend the format and also include other types of molecules. This has been mainly developed to support exchange of data rather than efficient storage.

The graphical representation of the XML schema structure of the PSI MI format is shown in (Fig. 3). Sources are used to describe the data sets that can be combined from various datasources. Experimentalist references to the experiments verifying the interaction. Interactorlist contains list of proteins participating in the interactions. Interactionlist contains a list of actual interactions. DIP supports two kinds of format (XIN and PSI MI) for exporting data, as shown in Fig. 4.

BIOPAX—Biological pathway data exchange) [32] format was designed to support metabolic pathway data and is implemented according to OWL standards which are the most important for biological pathways [33]. The BioCyc [17] BIND databases are committed to support this standard. BioPAX addresses issues related to the ontology used for representation of pathways. Different issues of integration concerning data types, external references and synonyms etc. are being addressed in BioPAX development. Entity is the base level and interactions, pathway and physicalEntity are sub levels as shown in (Fig. 5.). The entity also incorporates synonyms, comments and other references. It also contains utilityclasses that provides a custom data type when a simple data type, such as string or an integer is insufficient. The entity contains discrete biological units that describe pathways. The pathway consists of set of interactions. A pathway in BioPAX is a series of molecular interactions and reactions. It is proposed that metabolic and signal transduction pathways

would be subordinate levels of pathways. The Interaction subclass defines a single biochemical interaction between two or more entities. The main use of BioPAX level 1 is representation of metabolic pathways in the current state.

GeneXML—is one of the resources in GeneX project [34]. It is a specification that supports logical representation of data so that partial or complete data sets from different gene expression databases can be exchanged without information being lost. The GeneXML contains Entity as a main class and the entity contains Element and Attribute as subordinate levels as shown in Fig.6. The element can also be an entity which is divided in to further specifications. Gene expression data are generated by variety of sources that are stored in different levels of GeneXML format.

SBML—The SBML is a format developed by Systems Biology Workbench Development group (www.sbml.org) [35] that aims to serve as a future standard for exchanging mathematical models of molecular pathways. There are already a number of systems supporting SBML (e.g., JDesigner and Gepasi) mainly focusing on modeling and simulation. A pathway can be described as a model and each model can contain Compartments, Species, Reactions and Units respectively as main levels as shown in (Fig. 7). Compartment is a container where reactions take place. Species are the substances that take part in reaction. Reactions are the processes that modulate one or more species. Units are the descriptions of discrete changes in the model. In addition, a model can also contain definitions of parameters, mathematical functions, units and mathematical expressions. These are defined at the top level and can be used when defining the other entities.

CELL ML—(Cell Markup Language) [36] is an open standard based on XML that allows scientists to share models even if they are using different model-building software. CellML includes mathematics and metadata by leveraging existing languages that includes MathML and RDFs. The model is the root element for the CELLML document. It is the container for the components, connections, units and metadata. The geometric inclusions are handled at the component level. The CELLML structure is represented through a connectivity diagram that shows the different levels of the document as shown in (Fig. 8).

KGML—(KEGG Markup Language) [37] is the export format of the dataset from KEGG database, containing information about the exported KEGG pathway graph objects. The main level in KGML is Pathways and it has sub levels like entry, relation and reaction respectively. The entry level contains information about the individual graphs of pathways in the KEGG database. The relation specifies the relationship between two genes in the KEGG graph. The reaction specifies the chemical reaction between substrates and products.

3.2 DATA FORMATS FOR CHEMINFORMATICS

Many data formats exist for the representation of chemical structures. In general these representation formats vary in terms of descriptive richness. Chemical structures naturally occur as a three-dimensional configuration of atoms connected by bonds, but are often written as two-dimensional connectivity diagrams, or even one dimension strings such as a

chemical formula. This subsection reviews common chemical representation formats with attention to the level of information provided.

InChI—(The IUPAC International Chemical Identifier) [38] is a one-dimensional string representation of chemicals, developed by the IUPAC and NIST as a worldwide, human-readable standard for molecular representation. Unique InChI identifiers are generated from chemical structures by first normalizing the structure to eliminate redundant information, then unique indices are assigned to atoms and bonds, and finally converted into a serial string representation. The final string representation contains several “layers” of information such as chemical formula and connectivity, positive/negative atomic charge, stereochemistry, and isotopes.

SMILES—(The Simplified Molecular Line Input Specifications) [39] format is a one-dimensional string representation, constructed in such a way that the two-dimensional structure of a chemical can be recovered from the one-dimensional string. This is possible because the atom letters in a SMILES string are ordered according to a depth-first traversal of the chemical connectivity graph. Additionally, SMILES strings can encode a number of structural and physical properties of chemicals such as aromaticity, branching, stereochemistry and isotopes.

SMARTS—(The Smiles Arbitrary Target Specification) [40] is a string-based representation of chemicals that extends the SMILES format. The focus of SMARTS is on substructure specification a modifications to the representation language that allow SMARTS strings to function not just as molecular encodings, but also as queries that can be performed against chemical databases. These queries are not performed on the string representations directly, but rather the SMARTS/SMILES strings are converted to connectivity graphs and subgraph isomorphism is used as a criteria for matching.

MDL Molfile—is a widely used representation (named after the now defunct company MDL Information Systems that developed the format) that can encode structural information about a chemical compound [41]. This representation forms the structural basis for the better known Structured Data Format (SDF) file which contains MDL definitions for many appended compounds, along with additional name/value pairs of arbitrary tagged information (e.g., physicochemical properties or metadata) available on a per-compound basis. Along with a chemical identifier and header information, the MDL format lists a set of atoms and bonds occurring in a particular chemical, with two or three dimensional positions attached to the atoms, and connectivity information attached to the bonds. The SDF format is very flexible and most software systems that utilize chemical data will accept this format for import/export.

3.3. UNDERSTANDING FORMAT AND DATA SOURCE DIFFERENCES

The SBML and PSI MI are defined with an aim of being standards that can be used for exchanging pathway data. There also exist various tools that support these standards. The representation of interactions in all formats has at least one entity for representing subjects. The PSI MI specifies which molecule participates in interaction. This concept does not exist

in SBML, BioPAX and others, but there are some considerations being made. Representing reactions in these formats has been handled in different ways. SBML has several subtypes for representing reactions, whereas PSI MI, BioPAX and KGML also have only one way of representing reactions between interactors.

The principal structure of all formats is similar in that they reflect the structure of a pathway graph, such that pathway information is structured in representation of the interacting subjects. SBML represents the mathematical description of pathways, while the PSI MI format allows cross-references and inclusion of interactions concerning pathways, and BioPAX focuses on molecule interactions in metabolic pathways. Most of formats other than SBML or CellML support the addition of references to other databases in order to combine biological knowledge, however SBML or CellML are fairly unique in trying to store or represent mathematical knowledge required for understanding pathways. Hence both the biological and modeling knowledge are of high importance because they help to identify important pathway functionalities. The common issue that differentiates between these formats is the availability of a single format that could combine both the biological and mathematical modeling data into one syntax that incorporates biological knowledge based on the PSI MI format and stores mathematical descriptions required for simulations similar to that of SBML format.

Anticipating the future of XML as data exchange format for the biological databases particularly within pathway, protein interaction and mathematical modeling area, XML is a natural choice as a medium within which to encode such a unified syntax. Considering that BIND, DIP and PPI provide the protein interaction data sets in XML format based on PSI MI standards we integrated it into our proposed DMSPML representation, as illustrated in figure 9. Unfortunately KEGG is focused on the graphical presentation of pathway diagrams rather than biological concepts as in BIND or DIP, and thus its incumbent knowledge will be more challenging to integrate into a common exchange format.

The incorporation of mathematical modeling knowledge requires conducting specific studies for various pathway models, and at this point we are unaware of any quantitative databases that fully address the requirements for this purpose. Thus in order to integrate such knowledge into DMSPML, the mathematical models were exported in a standard form for simulation and visualization.

The chemical representations used by the previously discussed databases vary with the richness and amount/type of information used to encode chemicals. Rich, generalized formats such as MDL/SDF have been developed to represent molecules as 2D/3D atom coordinates and bond connectivities, along with an arbitrary amount of molecular descriptors or meta data attached as name/value pairs. This format has been widely accepted, allowing greater interoperability of chemical analysis software. The tagged nature of non-structural data and metadata in the SDF format could be readily converted to an XML-based representation if one wished to extend the global data exchange model to more fully recognize the informational content of chemical entities. It is uncertain whether the full extent of the structural detail embedded in SDF files (e.g., actual atomic Cartesian coordinates) is required for a common exchange, thus more compact, text string

representation formats such as the InChI, SMILES and SMARTS representations may be considered, especially considering their suitability as targets for database queries and chemical filtering mechanisms.

4. CONCLUSIONS

Today there are still many problems remaining in applying the combined biological and mathematical knowledge towards understanding the dynamic nature of pathways, and integrating the full wealth of chemical knowledge into the framework as would be critical for support of chemical biology and drug discovery. This study has assessed online databases, published mathematical pathway knowledge, and evaluates the basic nature of pathways using an integrative systematic approach. This systematic pathway analysis still has several limitations and further research must be pursued in order to produce results which can provide tangible help in finding possible targets for drugs.

This paper has reviewed a number of databases describing biological entities such as chemicals and proteins, along with possible relationships between them such as interactions, reactions, and pathways. Many of these databases contain tools for analysis and visualization of entities/relationships. These tools are especially useful for probing small molecule interactions, and a selection of database initiatives have arisen to directly facilitate drug discovery and development. These databases are funded by a variety of organizations around the world and while they share many features they are often designed for slightly different goals. Databases commonly vary according to several respects such as size, level of curation/annotation, or integration with other data sources, thus clearly a common standard for minimum information content and quality remains elusive. However the sheer volume of data currently available largely guarantees that a reasonable subset of the available knowledge has been curated with sufficient quality standards and relevant metadata representation as to be amenable to collective aggregation into a coherent body of knowledge spanning all of the disciplines from genomics/proteomics through pathways to chemical biology and drug discovery. Thus, a common exchange format to support this aggregation is a potentially vital step toward formulating next generation resources that can be used to support future advances in biological research and challenging emerging applications such as personalized medicine.

REFERENCES

1. Visvanathan, M.; Lushington, GH. Data Integration Issues and Challenges in Systems Biology. In: Plant, C.; Böhm, C., editors. Database Technology for Life Sciences and Medicine. World Scientific Publishing. 2010. ISBN: 978-981-4307-70-3
2. Brazma A. Editorial. On the importance of Standardisation in Life Sciences. Bioinformatics. 2001; 17:113–114. [PubMed: 11238066]
3. C. PubMed. 2000. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/>
4. Markowitz, VM.; Chen, IA.; Kosky, AS.; Szeto, E. Facilities for exploring molecular biology databases on the web: A comparative study. 2001. [Online]. Available: <http://gizmo.lbl.gov/DMTOOLS/OPM/WebInt/WebInt.ps>
5. Chaudhri, AB.; Zicari, R. Experience Using the ODMG Standard in Bioinformatics Applications. In: Paton, NW., editor. Succeeding with Object Databases: A Practical Look at Today's Implementations with Java and XML. Wiley; 2001.

6. Hammer, J.; Garcia-Molina, H.; Ireland, K.; Papakonstantinou, Y.; Ullman, JD.; Widom, J. Information translation, mediation, and mosaic-based browsing in the tsimmi system; Proceedings of the ACM SIGMOD International Conference Management of Data; 1995.
7. Chaudhri, AB.; Zicari, R. An Object-Oriented Database for Managing Genetic Sequences. In: Bellahsene, Z.; Ripoche, H., editors. Succeeding with Object Databases: A Practical Look at Today's Implementations with Java and XML. Wiley; 2001.
8. Fuentes G, Oyarzabal J, Rojas AM. Databases of protein-protein interactions and their use in drug discovery. Current Opinion in Drug Discovery & Development. 2009; 12(3):358–366. [PubMed: 19396737]
9. Bader GD, Christopher WV. BIND-a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. Bioinformatics. 2000; 16:465–477. [PubMed: 10871269]
10. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stümpflen V, Mewes HW, Ruepp A, Frishman D. The MIPS mammalian protein-protein interaction database. Bioinformatics. 2005; 21(6):832–834. [PubMed: 15531608]
11. Xenarios I, Salwinski L, Duan J, Higney P, Kim S, Eisenberg D. DIP: The database of interacting proteins. a search tool for studying cellular networks of protein interactions. Nucleic Acid Res. 2002; 30:303–305. [PubMed: 11752321]
12. Güldener U, Münsterkötter M, Kastenmüller G, Strack N, van Helden J, Lemer C, Richelles J, Wodak SJ, Garcia-Martinez J, Perez-Ortin JE, Michael H, Kaps A, Talla E, Dujon B, Andre B, Souciet JL, De Montigny J, Bon E, Gaillardin C, Mewes HW. CYGD: the Comprehensive Yeast Genome Database. Nucleic Acids Research. 2005 Jan 1; 33(Database issue):D364–D368. [PubMed: 15608217]
13. Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S. Expansion of the biocyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Res. 2005; 19:6083–6089. [PubMed: 16246909]
14. Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G. MINT, the molecular interaction database: 2009 update. Nucleic Acids Res. 2010 Jan; 38(Database issue):D532–D539. Epub 2009 Nov 6. [PubMed: 19897547]
15. D'Eustachio P. Reactome knowledgebase of human biological pathways and processes. Methods Mol Biol. 2011; 694:49–61. [PubMed: 21082427]
16. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004; 32:277–280.
17. Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, Lopez-Bigas N. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Research. 2005; 19:6083–6089. [PubMed: 16246909]
18. Nicolas Le N, Benjamin B, Alexander B, Melanie C, Marco D, Harish D, Lu L, Herbert S, Maria S, Bruce L, Jacky S, Michael H. Biomodels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Res. 2006; 34:D689–D691. [PubMed: 16381960]
19. Bhalla US. Database of quantitative cellular signaling. 2002 [Online]. Available: <http://doqcs.ncbs.res.in/>.
20. Sivakumaran S, Hariharaputran S, Mishra J, Bhalla US. The database of quantitative cellular signaling: management and analysis of chemical kinetic models of signaling networks. Bioinformatics. 2003; 19:408–415. [PubMed: 12584128]
21. The UniProt Consortium. The universal protein resource (UniProt). Nucleic Acids Res. 2008; 36:D190–D195. [PubMed: 18045787]
22. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res. 2009 Jul 1; 37(Web Server issue):W623–W633. Epub 2009 Jun 4. [PubMed: 19498078]
23. Seiler KP, George GA, Happ MP, Bodycombe NE, Carrinski HA, Norton S, Brudz S, Sullivan JP, Muhlich J, Serrano M, Ferraiolo P, Tolliday NJ, Schreiber SL, Clemons PA. ChemBank: a small-molecule screening and cheminformatics resource database. Nucleic Acids Res. 2008 Jan; 36(Database issue):D351–D539. Epub 2007 Oct 18. [PubMed: 17947324]

24. de Matos P, Alcántara R, Dekker A, Ennis M, Hastings J, Haug K, Spiteri I, Turner S, Steinbeck C. Chemical entities of biological interest: an update. *Nucleic Acids Res.* 2010; 38(suppl 1):D249–D254. [PubMed: 19854951]
25. Chen, Jonathan; Swamidass, S Joshua; Dou, Yimeng; Bruand, Jocelyne; Baldi, Pierre. ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics.* 21(22):4133–4139. [PubMed: 16174682]
26. Kuhn, Michael; Szklarczyk, Damian; Franceschini, Andrea; Campillos, Monica; von Mering, Christian; Jensen, Lars Juhl; Beyer, Andreas; Bork, Peer. STITCH 2: an interaction network database for small molecules and proteins. *Nucl. Acids Res.* 2009 Nov 6.
27. Li, Liwei; Bum-Erdene, Khuchtumur; Baenziger, Peter H.; Rosen, Joshua J.; Hemmert, Jamison R.; Nellis, Joy A.; Pierce, Marlon E.; Meroueh, Samy O. BioDrugScreen: a computational drug design resource for ranking molecules docked to the human proteome. *Nucleic Acids Research.* 2009
28. Bernstein F, Koetzle F, Williams B, Meyer J, Brice M, Rodgers J, Kennard O, Shimanouchi T, Tasumi M. The protein data bank: a computer based archival files for macromolecular structure. *Journal of Molecular Biology.* 1977; 122:535–542. [PubMed: 875032]
29. 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 AC, Wishart DS. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* 2011 Jan; 39(Database issue):D1035–D1041. [PubMed: 21059682]
30. Zhu F, Han BC, Pankaj Kumar, Liu XH, Ma XH, Wei XN, Huang L, Guo YF, Han LY, Zheng CJ, Chen YZ. Update of TTD: Therapeutic Target Database. *Nucleic Acids Res.* 2009
31. Yakugaku, Zasshi. KiBank: a database for computer-aided drug design based on protein-chemical interaction analysis. *Journal of the Pharmaceutical Society of Japan.* 124(9):613–619.
32. The BioPAX Consortium. The BioPAX Ontology Class Structure. 2004 www.biopax.org.
33. OWL Web Ontology Language. 2000 <http://www.w3.org/TR/owl-features/>.
34. GeneX Project collection. <http://lgmb.fmrp.usp.br/genex/>.
35. Finney A, Hucka M. Systems biology markup language (sbml) level2: Structures and facilities for model definitions. 2003 [Online]. Available: <http://sbml.org/documents>.
36. Catherine M, Lloyd M, Halstead B, Nielsen P. Cellml: its future, present and past. *Prog Biophy Molr Bio.* 2004; 85:433–450.
37. Kyoto University Bioinformatic Center. KEGG Markup Language Manual. <http://www.genome.jp/kegg/docs/xml/>.
38. McNaught, Alan. The IUPAC International Chemical Identifier: InChI. Chemistry International (IUPAC). 2006; 28(6)
39. Anderson, E.; Veith, GD.; Weininger, D. SMILES: A line notation and computerized interpreter for chemical structures. Report No. EPA/600/M-87/021. Duluth, MN 55804: U.S. EPA, Environmental Research Laboratory-Duluth; 1987.
40. SMARTS Theory Manual, Daylight Chemical Information Systems. Santa Fe, New Mexico: <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>
41. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, et al. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of Chemical Information and Computer Sciences.* 1992; 32:244–255.

Figure 1.
Interdisciplinary relationships between fields related to bioinformatics.

Figure 2.

Biological entities and some relationships between them. Chemicals and small molecules may interact with or inhibit proteins as well as DNA or RNA. DNA is related to RNA through transcription, and RNA are related to proteins through translation. Proteins may in turn interact with other proteins or regulate certain genes. Many kinds of biological entities participate in pathway reactions or binding events.

Figure 3.
Graphical representation of the PSI MI format structure.

Figure 4.
DIP Database exchange format based on PSI MI standards.

Figure 5.
Graphical representation of BioPAX structure.

Figure 6.
Representation of the GeneXML structure.

Figure 7.

A graphical representation of SBML document structure.

Figure 8.
Graphical representation of CELLML document structure.

Figure 9.
The concept behind the new proposed export format of DMSPML.

Table 1

A listing of categories for database review, along with typical (although not exclusive) database entities and some example databases.

Category	Typical Entities	Example DBs
Protein Interactions	Proteins, DNA, RNA, interactions	MIPS, DIP, CYGD, EcoCyc, MINT
Pathways	Proteins, chemicals, reactions	REACTOME, KEGG Pathway, BioCyc
Mathematical Modeling	Quantitative pathway models	Biomodels, DOQCS, UniPathway
Chemical Interactions	Chemicals, proteins, interactions	PubChem, ChemBank, ChEBI, ChemDB, STITCH
Drug Discovery	Chemicals, proteins, activities, interactions, reactions	BioDrugScreen, DrugBank, TTD, KiBank, SMID