Published in final edited form as:

Curr Protoc Chem Biol. 2012 September ; 4: 193–209. doi:10.1002/9780470559277.ch110262.

Dealing with the Data Deluge: Handling the Multitude Of Chemical Biology Data Sources

Rajarshi Guha*, Dac-Trung Nguyen, Noel Southall, and Ajit Jadhav

NIH Center for Advancing Translational Science, 9800 Medical Center Drive Rockville, MD 20850

Abstract

Over the last 20 years, there has been an explosion in the amount and type of biological and chemical data that has been made publicly available in a variety of online databases. While this means that vast amounts of information can be found online, there is no guarantee that it can be found easily (or at all). A scientist searching for a specific piece of information is faced with a daunting task - many databases have overlapping content, use their own identifiers and, in some cases, have arcane and unintuitive user interfaces. In this overview, a variety of well known data sources for chemical and biological information are highlighted, focusing on those most useful for chemical biology research. The issue of using multiple data sources together and the associated problems such as identifier disambiguation are highlighted. A brief discussion is then provided on Tripod, a recently developed platform that supports the integration of arbitrary data sources, providing users a simple interface to search across a federated collection of resources.

1 Data Sources in the Life Sciences

With the development of modern molecular biology and high throughput techniques, the amount of chemistry and biology data generated from experiments has exploded. Depending on the nature of the experiment and technique, the data generated may be directly interpretable (e.g., dose-response curves from high throughput titration experiments) or may require significant postprocessing to transform it into a meaningful form (e.g., next generation sequencing data). In either case, the resultant data is usually deposited in a public database and may or may not include metadata describing the data and how it was generated and processed. The scale and breadth of modern experimental chemistry and biology has led to a profusion of publicly available databases containing a wide variety of data. Most databases are domain-specific, although many make significant efforts to integrate many different types of data. Examples include Pubchem and ChEMSPIDER, whose primary data is chemical structure. However, both databases make extensive links between chemical structures and other relevant data types including biological activities, spectra, protein targets and ADMET properties. Furthermore, the size of databases varies widely. For example, Pubchem contains more than 28 million entries whereas Drugbank has less than 10,000 entries. Indeed, there are many databases with just a few hundred entries. A characteristic feature of some databases is the use of manual curation, where humans

* Author for correspondence: NCTT/NCATS - Informatics, 9800 Medical Center Drive, Rockville, MD 20850, Tel: 301 217 2337, guhar@mail.nih.gov.

manually collect the data or examine the deposited data to ensure its correctness. While manual collection is usually associated with smaller databases, this is not necessarily always the case. As an example, the ChIP-X database published by Lachmann et al (Lachmann, Xu et al. 2010) manually analyzed 87 publications to extract 189,933 interactions between transcription factors and genes. At the other end of the scale, the ChemSpider database contains records for more than 26 million chemical compounds. This database also allows the community to curate records, and thus while not a manually created database, over time it is expected to become a manually curated database.

2 Databases for Chemical Biology

Many biological databases are of general interest, covering genes, proteins, micro RNA's, either across species or in a species-specific manner. While these are applicable to problems in chemical biology, databases that consider small molecules and their biological activities are of special interest to this community. In addition, there are a number of databases that link small molecules to diseases and phenotypes, via target information. Even in cases where such links are not explicit, such databases can be of use with additional computational analysis. While many databases focus on chemical structures, the real value arises when they are linked to reports of biological activity, thus providing information on structure-activity relationships (SAR). The following sections highlight a number of databases in different topical areas, all relevant to problems in chemical biology.

2.1 Chemical structure databases

Databases of chemical structure represent the most relevant source of information in chemical biology datascources. As noted above, databases that link chemical structure to biological activity support SAR analyses and so go beyond simple look up of chemical structures. In general, few chemical structure databases are pure repositories of structures. Most such databases will annotate structures with a variety of information ranging from physical properties to literature and patent references. Table 1 lists a number of chemical databases based on a listing of 64 free chemistry databases by Apodaca (Apodaca 2011), highlighting those that provide SAR information. Of particular utility are the ChEMBL and GOSTAR databases. The former, provided freely by the EBI and the latter provided under a commercial license from GVK Biosciences, focus on chemical structures and their bioactivities in a wide variety of screens, taken from a large number of peer-reviewed publications. Both are manually curated, thus providing a degree of reliability. In both cases, the database providers have taken extra steps to either provide extensive annotations (such as assay target, species and cell line) or link into other domain specific data sources, such as toxicity and PK/PD data for the case of GOSTAR. In both cases, each activity record is linked to the original publication. On the purely structural side, databases such as SciFinder provide a vast amount of information, curated from the chemical literature. This database does not link into bioactivity data sources, but provides links to compound vendors as well as extensive information on physical properties including spectra as well as linking to synthetic protocols and so on. Structure databases that link into spectral (NMR, infrared, mass) information are also useful, especially in structure elucidation problems. Both free (Sigma-Aldrich, Spectral Database System, NMRShiftDb (Steinbeck and Kuhn 2004)) and

commercial (Reaxys, SciFinder, BioRad) databases are available. Other, smaller databases focus on specific subsets of small molecules. For example, DrugBank (Wishart, Knox et al. 2006; Wishart, Knox et al. 2008) considers drugs and integrates small molecule structure information with extensive annotations on the targets of the drugs, dosage, side effects and interactions and so on. The KEGG database (Ogata, Goto et al. 1999) focuses on small molecules and drugs that modulate metabolic pathways.

A key feature of chemical databases is the ability to perform chemical structure searches in a variety of ways. For example, exact structure searches are performed by specifying SMILES (a linear, text based format for the representation of chemical structures) strings or uploading MOL (a multi-line text based format that is capable of encoding 3D representations of chemical structures, along with a variety of associated information such as R-groups) files. Another common search strategy is to identify compounds that contain a specific substructure or identify structures that are “similar” to a query structure. The latter can be challenging as there are many ways to define similarity, ranging from fingerprint similarity (focusing on connectivity and presence/absence of functional groups) to shape similarity. Exact and substructure searches are generally very fast. The speed of similarity searches depends on the degree of similarity specified. For very high cutoffs (i.e., only return very similar molecules), the search can be very fast. As increasingly dissimilar molecules are desired, the search time increases. It should be noted that most online databases only support the former type of similarity that only considers connectivity and functional groups. Shape searching is not very common, primarily due to the need to store large numbers of conformations to ensure comprehensive coverage of the shape space. Nearly all chemistry databases support a variety of inputs, including uploading MOL files, pasting SMILES strings and drawing structure diagrams.

2.2 Patent databases

Patent databases play an important role in the day-to-day scientific work. A common task is to identify prior art, before embarking on a screening or synthetic program and patents are a primary resource for such information. Most patent authorities make their patent documents freely available. Chemical structures are usually represented as images, though more recent patents do contain MOL files of the structures mentioned in the document. However, these are usually generated post hoc and therefore, in many, cases require extensive cleanup. Furthermore, patent documents may refer to chemicals via a numbering scheme or even by common names, all of which must be linked to actual structures. Finally, most patents will employ Markush structures, general patterns defining an entire class of compounds. These structures are generic in nature, usually defined in terms of an explicit “core” structure that is present in all the compounds defined by the pattern, along with a variable number of groups attached at (possibly) different positions on the core. Most tools that are available to represent and manipulate these types of structures are commercial in nature. Because of the effort needed to maintain high quality structure information from patents, most patent databases are commercial in nature, requiring licenses or subscriptions for access. Examples include SciFinder STN, Reaxys, Derwent, Questel and so on. Patent databases that are free (which are usually maintained by the patent authority) usually allow only text searches and do not provide any chemical intelligence but can be helpful for identifying chemical matter

that interacts with a specific target. ChemSpider, which is free, provides some indexing of patent matter and links back to the original document via its collaboration with SureChem. PubChem has many patent compounds indexed, but linking back to the actual patent document requires a subscription with the original depositor.

2.3 Disease & phenotype databases

Much of chemical biology research addresses therapeutic or potentially therapeutic small molecules for a variety of diseases. Such research benefits from a variety of information on diseases including molecular basis, known treatments and various classifications of diseases. OMIM is an extensive collection of phenotypic information with their corresponding genotypic associations. In addition, a variety of more specific disease databases have appeared covering wider classes such as autoimmune diseases (Karopka, Fluck et al. 2006) and rare diseases (<http://www.orpha.net/> and <http://www.rarediseases.org/rare-disease-information/rare-diseases>). A notable resource for drug, gene and disease information is PharmGKB (Klein, Chang et al. 2001; Thorn, Klein et al. 2005) that focuses on the effects of genetic variations on drug responses and integrates a variety of resources including pathways, clinical trial data, genotypic and phenotypic information. Clinical trials can be a useful source of drug and disease information and the primary source for this type of data is at ClinicalTrials.gov which supports full text search as well as download of the data. Note that much of the information in this data source is in the form of free text and small molecule information is generally available in the form of common names, rather than IUPAC nomenclature or formats such as SMILES. An alternative view of this area is provided by databases that describe the effects of small drugs in terms of symptoms. Examples include the SIDER database (Kuhn, Campillos et al. 2010) which captures side effect information of drugs and drug pairs.

While the above data sources are a rich source of information connecting molecular level data to phenotypic data, automated classification and unique identification of diseases is still a significant challenge. Identifiers play a significant role in querying across databases and merging different databases (Section 4). The SNOMED CT is a nomenclature system that broadly covers health care concepts such as myocardial infarction or viral pneumonia. It is hierarchical in nature, containing high level categorizations (such as disease classes) as well as instances of specific diseases. Other approaches to classification are the ICD (International Classification of Disease) codes and MeSH terms that assign unique identifiers to disease and disease concepts. More recently the Disease Ontology (Schriml, Arze et al. 2011) attempts to integrate the multiple classification systems into a formal ontology that would allow automated methods to analyze relationships between diseases. Many of the disease databases listed above make use of one or other of the classification systems to link diseases to genes and small molecules. However, it is still the case that many resources will use common names (and variations on them) to refer to diseases. As a result, some form of postprocessing must be performed to uniquely identify the disease being referred to and this is an ongoing area of research (Osborne, Flatow et al. 2009; Wall, Pivovarov et al. 2010).

2.4 Protein & target databases

Many chemical biology studies focus on specific targets or families of targets. In such cases, where the target is known, structure-based modeling methods such as docking and pharmacophore modeling can provide insight into modes of action and lead optimization efforts. In these cases crystal structures of the target are vital. Protein structure information is available in a number of databases. The Protein Data Bank (PDB) is probably the best known of these databases and currently has records for more than 18,000 protein structures at a variety of resolutions, many of them in their ligand bound forms. The PDB has given rise to a number of derivative, focused databases. An example is PDBind (Wang, Fang et al. 2004) which collects PDB records for which there are experimentally measured affinity data. The database is very useful for developing virtual screening methods and benchmarking new docking algorithms. UniProt is key protein target database, but focuses on sequence rather than structure. It is an important reference database, as many other databases will refer to protein targets via their UniProt identifiers. Other sequence oriented databases include the NCBI Protein resource, the Protein Information Resource (PIR) (Wu, Yeh et al. 2003) and various resources from the EBI (<http://www.ebi.ac.uk/Databases/protein.html>). The Therapeutic Target Database (TTD) is collection of 2,025 targets and 17,816 small molecules associated with those targets. As the name suggests the databases focuses on targets that are of therapeutic interest. The collection also provides a variety of annotations related to target validation, QSAR models, as well as mechanisms of action for many of the small molecules. A somewhat similar database is the Potential Drug Target Database (PDTD) (Gao, Li et al. 2008) that integrates 3D protein structures that are known (and potential) drug targets with other data sources such as the TTD and DrugBank. A number of protein structure resources focus on protein families and domains such as Pfam and SCOP (Murzin, Brenner et al. 1995). Knowledge of domains can be useful in chemical biology applications as they can provide insight into selectivity (or promiscuity) of structural motifs as well as guiding library design (Dekker, Koch et al. 2005).

The mechanism of action (MoA) is a valuable piece of information for small molecules. However, there is no single (freely available) database that provides a comprehensive list of mechanism of actions that is tied to specific protein targets or genes. A commercial resource for this information is the Prous Integrity database. While the database provides MoA information for many small molecules, it is in the form of free text and may or may not include a gene symbol or protein target name. As a result, it is useful to apply a variety of text mining methods to integrate this information with other resources such as interaction and pathway databases. Other resources for MoA information are DrugBank and PharmGKB, though as with Integrity, this information is in the form of free text. In addition to these resources, smaller, more focused collections also exist. An example is the NLM Dietary Supplements Label Database (<http://dietarysupplements.nlm.nih.gov/>) which collects information on dietary supplements and in many cases includes MoA information for individual components. One of the challenges in formalizing MoA information is the variety of effects that a small molecule can have against various targets. For example, the MoA of ephedrine derives from both its direct and indirect effects on the adrenergic receptor system.

2.5 Pathway & interaction databases

To obtain a systems level view of small molecule activity (Oprea, Tropsha et al. 2007) it is important to have data on pathways that small molecules are involved in, interactions between proteins that might be disrupted by small molecule interventions, as well as kinetic data when possible. Table 2 lists a number of databases focusing on pathways and interactions and many of the problems discussed in Section 3 also apply to these databases. Thus, it can be a challenge to map small molecules from KEGG to GeneGO. More worryingly, pathway definitions can differ from one source to another as identically named pathways can contain slightly different sets of proteins. Soh et al (Soh, Dong et al. 2010) reported that the consistency of genes in similar pathways ranges from 0% to 88% across three different pathway databases. A number of vendors have collected and curated multiple pathway databases, as well as developed their own pathway maps. Examples include Metacore from GeneGo and IPA from Ingenuity Systems. The original problem certainly remains, in that the same pathway (identified by name) from two different databases may not be identical in terms of nodes and connectivity. The advantage of these commercial products is that they tend to be quite comprehensive as they combine not only pathway information but also gene, target, small molecule and literature resources into one integrated system.

3 Challenges in Dealing with Multiple Databases

Clearly, there is a plethora of information available for the practicing scientist. While a number of resources are commercial, many resources are freely accessible. However, this multitude of data sources is associated with a number of problems and challenges, many of which are not yet solved. Philippi et al (Philippi and Kohler 2006; Philippi 2008) provide a review of these problems and this section focuses on a few of them.

Probably the biggest challenge faced by experimental biologists and chemists is to determine which database is useful for their needs. In many cases the first stop is the collection of databases provided by the National Library of Medicine (e.g., PubMed, Entrez Gene, PubChem). These are especially useful since many of these are linked to each other. Thus a search for a compound in PubChem can lead directly to the NCBI Protein database, if the compound has been annotated with a protein target. Section 3.2 discusses some uses cases from chemical biology and the relevant databases that can be employed to address them.

However, there are many other databases, both commercial and public that can provide different, but complementary information. A useful source of information for new databases is the special database issue of Nucleic Acids Research. Figure 1 summarizes the number of new databases reported in the yearly special issue over time. The Nucleic Acids Research online Database Collection, (<http://www.oxfordjournals.org/nar/database/a/>), currently lists 1,338 molecular biology databases.

Beyond the increasing number of available databases, users are faced with several other problems. One of the primary bottlenecks is the issue of identifiers. For example, the gene HSP90AA1 has 13 aliases within Entrez Gene. Furthermore, it is also reported in other

databases include Ensembl and UniProt, which use the identifiers ENSG00000080824 and P07900. While it is true that most large databases will store the multiple synonyms and identifiers from other databases, there is no guarantee that the coverage is complete. The problem manifests itself greater for chemical structures, due to the challenges associated with standardizing structures in a consistent fashion across databases and vendors. In some cases, automated methods are employed to impute links. An example is to apply text mining algorithms on the description of a clinical trial or a paper abstract, to identify disease names and then annotate the item with terms from a formal ontology or classification (such as ICD-10 or SNOMED). The annotations on different items can be used to link them together. Given that such approaches are rarely 100% correct, invalid annotations can be assigned and as a result invalid or nonsensical links can be generated. While useful to obtain an initial set of links, high quality databases will employ some form of manual curation to ensure that the links are reasonable. More fundamental than identifiers is the fact that different databases may differ in their internal representations of a concept. For example, one database may link a small molecule to an enzyme, simply indicating that the two are associated, whereas another will report the association but also include the observed stoichiometry. Even then, variants of an enzyme can catalyze different reactions and hence the specific variant should be included to fully specify the combination of small molecule, enzyme and reaction (Philippi and Kohler 2006).

Another issue with the large number of databases is that many (smaller) databases are derived from other databases. Consequently, errors in the source database can be propagated into the downstream databases. In some cases, corrections in the downstream database are fed back to the upstream databases; but this is not common practice. This challenge is closely related to the issue of curation in databases. While some smaller databases are generated manually (and thus curated), most databases are not. This means that when errors are discovered there may not be a way for a user to suggest a correction (beyond sending a report to the database maintainer). Of course, for large collections such as PubChem, curation of the entire database is not feasible. However some databases such as ChemSpider and the NCGC Pharmaceutical Collection (Huang, Southall et al. 2011) do provide curation interfaces, that allow users to directly provide a correction, which is then merged back into the database.

A major challenge facing users who simply want to find an item of interest is the myriad ways database designers create user interfaces. While some larger organizations such as NCBI provide a single interface to all their databases, it is still a different interface than that created by the EBI. Within any given interface, there can be many options to refine searches or define searches in different ways (such as specifying a SMILES string or drawing a structure when searching for compounds). In effect, a user must spend some time learning the interface of every new database that they wish to use.

The problem of different, incompatible interfaces is also evident at lower, technical levels. For example, there are scenarios queries need to be performed in bulk (e.g., characterize the promiscuity of fifty compounds identified from a HTS campaign). While this could be performed manually, it is much more efficient to address this in a programmatic manner. Yet, even for a programmer, each database presents a different interface. Some databases

provide a simpler interface (e.g., REST), whereas others are more complex and in many cases, there is no programmatic interface at all. In this last scenario, batch queries can only be realistically performed by rehosting the entire database locally. Even then, there are cases when no programmatic interface or data dumps are available and data must be accessed via the provided web interface. In this situation “scraping” the website is the preferred approach. That is, programmatically access individual web pages and parse the page to extract the relevant information. As expected, this can be a trial and error process and in many cases violates terms of use of the database.

The preceding discussions have focused on technical problems faced by the life sciences databases. But the community also faces non-technical problems. For example, funding and long-term maintenance is an important parameter when choosing a database to use. In many cases, a database is not maintained over an extended period of time (usually only during the funding period). This can be seen for a number of smaller, academic databases. In the best case, the database server (or the actual data) is simply made available as a dump file, so that other interested parties can re-host them. In the worst case, the database simply disappears. This is generally not the case for commercial database vendors. Licensing and access restrictions can severely affect the integration process. Obviously, commercial databases will have restrictions on reuse and redistribution of content and in most cases must be treated as silos. However, even for free databases, the presence of onerous licensing policies or even the absence of any license or use policy can prevent automated analysis and integration efforts.

3.1 Searching for chemical structures

It is evident that the profusion of different identifiers referring to the same object (gene, protein, etc.) are a key source of problems when working with multiple databases. This problem is aggravated when linking one or more structural representations structure to an identifier. An incorrect approach to this problem can easily result in not finding a compound when searched for by name, SMILES or structure. The problem fundamentally arises due to the complexity of chemical representations. For example, should salts be stripped from a compound? Which tautomeric form should be stored in the database? Since a chemical structure can be represented using different SMILES, which one should be stored?

Most of these questions are addressed by a standardization algorithm. This algorithm will address several aspects including

- Strip salts from a structure
- Derive a canonical tautomeric form
- Generate a unique canonical SMILES from the given structure

The result of a standardization algorithm is that it will always generate a unique, canonical representation of that structure, irrespective of the input representation. Note that this applies to inputs that are structural representations, such as SMILES, MOL and SDF formats, as opposed to chemical names. However, different databases employ different standardization procedures. As a result, while it is reasonably true to assume that a given database has a unique representation of a given structure, there is no guarantee that this same

representation will be found in another database. This can make simple linking between chemical databases difficult.

One recent approach has been the use of the InChI string and InChI Key representations (McNaught 2006). These are canonical string representations of a chemical structure developed by IUPAC. While these representations are not human interpretable, the fact that they are defined by a specific source (until now) implies that every database using a InChI to represent their compounds can now directly link records without having to perform extra processing or standardization steps. Note that InChI is not infallible as cases have been discovered where different molecules have the same InChI. However these are usually very rare occurrences.

The InChI software is freely available, allowing anybody to generate their own InChIs. In addition services such as the NCI Chemical Structure Lookup Service (CSLS) and ChemSpider return a chemical structure given the InChI Key representation. The reader is referred to Warr (Warr 2011) for an extensive review of chemical structure representations.

Clearly, there are a number of steps involved in standardizing chemical structures such that they are searchable and comparable in a reliable fashion. Fourches et al (Fourches, Muratov et al. 2010) have discussed this issue and described computation pipelines to ensure correctness and integrity of chemical structure data. In addition, a number of tools to perform these standardization steps are also available. Examples include the Standardizer from ChemAxon and a standalone standardizer available from the authors (<http://tripod.nih.gov/ws/standardizer/standardizer.jnlp>).

Structure searching becomes more problematic when faced with chemical names. In some cases, such as when using formal IUPAC nomenclature, software can parse such names to obtain a structure (e.g., Lexichem from OpenEye and OSCAR4 (Jessop, Adams et al. 2011)). More frequently, however, is the use of common chemical names or trade names. Since these are arbitrary identifiers, the only way to “resolve” such a name to a structure is to look it up in a dictionary. If the dictionary does not have an entry for the given name, then no structure can be obtained. Clearly, the process of building such a dictionary is a never-ending process. In addition, variations in spelling and punctuation (dashes, commas etc.) must be taken into account. A number of data sources provide extensive coverage of chemical names. PubChem provides an extensive list of synonyms for PubChem Compound Ids (50,117,829 synonyms for 19,622,291 compounds). The NCI Chemical Identifier Resolver (<http://cactus.nci.nih.gov/chemical/structure>) is also a useful resource for name to structure lookups and provides a simple REST interface that allows bulk lookups in a straightforward manner. A useful feature of this service is that it is capable of approximate string matching (a technique that can match strings that differ only slightly from each other, such as by a single misspelt letter or single hyphen) and thus “diethyl fumarate” and “diethyl fumarate” are resolved to the same structure.

3.2 Use cases & relevant databases

This section highlights a number of common use cases from chemical biology and some databases that are applicable to the problems. This is neither an exhaustive list of use cases

nor of the applicable databases, but serves to highlight how multiple databases (or if sufficiently integrated, a single database) can be used to explore problems in chemical biology.

HTS hit triage—A key step during an HTS campaign is the selection of hits from the primary screen. Since a large HTS tends to identify a large number of hits, a triage procedure must be performed to select a manageable subset of those hits that will most likely reconfirm. A number of approaches have been described (Yan, Asatryan et al. 2005; Simmons, Kinney et al. 2008; Posner, Xi et al. 2009; Reitz, Smith et al. 2009). A common approach is to make selections at the scaffold level, rather than that of individual compounds. In other words, the set of scaffolds present in the screening collection is first characterized and then prioritized based on enrichment in active compounds, or any other scoring function that ranks scaffolds. Then, from each scaffold the requisite number of compounds are selected. Ideally, such analyses will make use of many data sources - how common is the scaffold in other HTS campaigns, how many times do compounds with this scaffold show activity and so on. While many databases such as ChEMBL, PubChem and GOSTAR provide small molecule -bioactivity associations, they do not provide scaffold information. As a result, users must generate their own scaffold data and create associations between scaffolds and the activity of the scaffold members. Section 5 describes how the Tripod software platform integrates this type of scaffold-based analysis with a variety of external databases.

Promiscuity analysis—When performing selections from a HTS campaign or choosing compounds for lead optimization, it is useful to determine the promiscuity of the compound. Promiscuous compounds can show up as actives in a screen due to a variety of reasons, not necessarily related to on-target activity. Thus it is useful to screen out such compounds. A simple measure of promiscuity is to identify the assays a compound is tested in and in how many of them it is identified as active. Two databases suitable for this are PubChem and ChEMBL. Both databases can be queried to identify the compound of interest and if present, identify the number of assays it has been tested in. Both PubChem and ChEMBL provide real-valued activity data (i.e., K_i , IC_{50} , etc.) and these values must be thresholded to classify compounds as active or inactive. PubChem does provide depositor supplied calls of active/inactive, thus avoiding the need to subjectively define a threshold. Recently, Canny et al have provided an online service that reports the promiscuity of a compound based on PubChem data (Canny, Cruz et al. 2011).

Target identification—In many cases, a phenotypic assay is designed without a specific target in mind. Thus, the specific target for hits in such assays are unknown and so target identification can be an important step in characterizing the mechanism of action. While accurate target identification requires experimental validations, databases can be employed to determine a set of possible targets and in some cases, the actual target. A useful resource for this type of task is ChEMBL, which associates assays with targets. Obviously, a compound that is active in an assay is not necessarily active against the target annotated to that assay. However, it is a reasonable starting point. ChEMBL can be queried to identify a specific compound and subsequently the assays it was tested in and the targets of those

assays. If the query is not present putative targets can be suggested based on the targets of compounds that are structurally similar to the query compound. As noted in Sections 2.3 and 2.4, disease and protein databases can also provide links between small molecules and their targets. DrugBank and PharmGKB for example, lists targets drugs, where known and KEGG can be a useful resource when working with metabolic pathways.

Quantifying novelty—A useful characteristic when characterizing targets is the “novelty” of the target. For example, a primary RNAi screen can identify tens to hundreds of hits. In addition to the usual metrics of signal strength, reproducibility across replicates and so on, it is sometimes useful to characterize the novelty of the selected genes. In this context, the term novelty refers to the popularity of the gene in the scientific literature (Su and Hogenesch 2007). Thus a gene that has been extensively published on in the literature may be very well known and thus not as interesting as a gene for which only a few papers are available. (Of course, the lack of literature around a gene does not imply that it is novel since it might be the case that the gene has been recently discovered or even that the gene is so boring that few people are willing to study and publish on it). A novelty score can be determined via a PubMed search, where the number of publications returned is used as a rough measure. A somewhat more quantitative approach is to directly use the Entrez gene id - Pubmed id associations that are available from <ftp://ftp.ncbi.nih.gov/gene/DATA/gene2pubmed.gz>. This data file lists genes and the associated documents in which they were mentioned. Given a gene id (which can be obtained for a given gene name via Entrez Gene), it is easy to identify how many papers it is associated with. Crudely, for a given gene, a large number of papers suggests lesser novelty. While it is quite simple to parse and process the data file, a simple web interface is available, (<http://assay.nih.gov/novelty.html>), that takes in a list of gene names and obtain the corresponding novelty scores.

Identifying interacting targets—In many cases, the protein target of a small molecule is known. However, to gain a broader picture it can be useful to understand how the target interacts with other proteins and thereby understand downstream effects of the modulation of the target protein by small molecules. These types of questions are best answered using protein interaction databases. Some of them such as HPRD (Peri, Navarro et al. 2003) focus on protein-protein interactions, whereas others such as STITCH (Kuhn, Szklarczyk et al. 2010) consider interactions between small molecules and proteins in addition to protein-protein interactions. The MIMI database (Tarcea, Weymouth et al. 2009) is especially useful as it has integrated multiple interaction databases, covering small molecules, proteins and genes, into a single interface that is closely tied to a variety of NCBI resources. On the commercial side, MetaCore from Genego and IP provide extensive interaction (and connectivity) information on proteins and small molecules, with easy to use interfaces.

Cross-assay analyses—Many chemical biology programs focus on the effects of small molecules on a specific target or pathway. As pointed out by Oprea et al (Oprea, Tropsha et al. 2007), systems chemical biology, which takes into account the effect of small molecules on larger biosystems (i.e., networks of proteins and pathways) can provide a much broader and mechanistic view of small molecule bioactivities. To achieve a systems level understanding of such bioactivity requires an analysis of small molecule behavior across

multiple targets. Databases such as PubChem and ChEMBL provide the data resources to perform such cross-assay analyses, though actually performing such analyses is not always easy. Performing such cross-assay analyses can be very simple such as developing predictive models for a set of small molecules against multiple targets (i.e., one model per target). More sophisticated approaches take into account connectivity between targets and use network modeling approaches to address polypharmacological effects of small molecules (Metz and Hajduk 2010). GeneGo is notable for its focus on network analytics, and provides a number of canned algorithms to characterize gene and protein networks as well draw inferences from a variety of experimental data in the context of known networks (Brennan, Nikolskya et al. 2009).

Tracking historical trends in approved drugs—Historical data on drugs can be an invaluable resource for a variety of reasons. Identifying trends in various properties of drugs over the years or identifying the most recently approved drugs in a certain class can provide valuable information in a drug development project. The Drugs@FDA database, (<http://www.accessdata.fda.gov/scripts/cder/drugsatfda/>), is a resource for information on approved drugs in the United States. The website allows users to search for drugs by name, active ingredient as well as browse approvals by month. Each drug is associated with FDA application number, approval date and so on.

The approval date from the DrugsFDA database combined with other databases can be used to investigate a variety of trends. For example, the trend in hERG liabilities over the last 15 years or so can be analyzed using this data. For example, by combining approval date with the hERGAPDbase (Hishigaki and Kuhara 2011) (or even running the molecules through a hERG liability predictor (Sun 2006; Kramer, Beck et al. 2008)), it is observed that industry has recently relaxed its constraints on compounds with hERG liabilities and submitted NDAs for a number of drugs with in vitro liabilities that are nevertheless safe and effective in vivo. Similarly, approval dates of drugs could be combined with information on their targets (via DrugBank and Uniprot) to summarize how different target families (GPCR's, kinases etc) have been addressed over time.

4 Integration vs. Federation

Given the large (and increasing) number of databases, how can the issue of different identifiers, interfaces and so on be addressed? One approach is to federate databases under a common interface. In this way, a single query can be used to interrogate multiple databases and the results returned at one go. An example of this approach is the Entrez cross-database search. Figure 2 highlights how the query term “HSP90” is searched for across multiple databases and the number of results in individual databases is listed for each one. One of the key features of a federated system is that it does not attempt to merge different databases into a single monolithic system. Thus, Entrez allows the user to go directly to any of the individual databases and employ its query interface directly. The utility of a federated system is that it performs this task (which will usually differ from database to database) in the background so that the user does not need to bother with individual databases (at least, initially).

An integrated system on the other will usually combine the data from multiple sources and store it in a single local system. This process usually involves mapping different (equivalent) identifiers to a standard form as well as linking items that have some relationship to each other. An example of such an integrated database is canSAR (Halling-Brown, Bulusu et al. 2011) which integrates a variety of data sources to provide a comprehensive collection of data related to cancer drug discovery. The integrated nature of the database is highlighted by the fact that it combines results from chemical screens, RNAi experiment, expression profiles and so on. By virtue of links made between individual data types, it is easy to jump from a target of interest to viewing expression levels in different tissues to known compounds that inhibit the target and so on.

In some cases, a combination of federation and integration are required. An example is the PathwayApi resource (Soh, Dong et al. 2010) that unifies the pathways from three sources: Wikipathways, Ingenuity and KEGG. In this case, the individual databases are not merged, but preprocessing steps were performed to correctly match up similar pathways from the different sources. As a result, the user can query a single interface to obtain the correct pathway from the three pathway databases considered.

It is apparent from this discussion that both federation and integration of multiple databases requires significant efforts. With the exception of larger organizations, most examples of these databases tend to be sub-field specific. Furthermore, it is usually not possible for a user to design a “custom” integrated (or federated) database that combines just the data sources that they are interested in.

4.1 Linking data semantically

Federation or integration are time-consuming processes and require significant manual processing to correctly map fields and terms from one database to another. The concepts underlying the semantic linking of data has steadily grown over the last ten years. The basis of this approach is that by addition of meta-data, objects (which can range from database records to web pages) can be linked to each other in such a way as to provide meaning or context. The idea of machine understandable descriptions of “things” is supported by a variety of technologies such as RDF, OWL, SPARQL and formalized descriptions of specific domains in the form of ontologies.

One immediate utility of these technologies is to generate links between data sources that are related but not identical. This approach makes use of an ontology that describes how objects, for a specific domain are related to each other. For example an ontology for chemistry would specify that a molecule is made of atoms, bonds contain two or more atoms and that a chemical substance can take on multiple forms. Annotating records with ontology terms allows a computer system to automatically identify fields or records in different databases that mean the same thing.

Another useful application is the ability to indicate that records refer to the same thing, even though the actual details of the records might be different in different databases. This makes merging databases, or querying across multiple databases much more efficient. An excellent example of the use of semantic technologies to create a network of linked resources is the

Linked Open Drug Data (LODD) effort (Samwald, Jentzsch et al. 2011) which has integrated twelve open access data sources, including DrugBank, ClinicalTrials.gov, ChEMBL and others into a collection of RDF triples, which are a representation of individual relationships between objects (such as a drug molecule and a clinical trial in which it is being tested). Other examples of semantic integration efforts include Bio2RDF (Belleau, Nolin et al. 2008) and Chem2Bio2RDF (Chen, Dong et al. 2010).

It should also be noted that data sources may provide their data in a form that is suitable for inclusion into semantics enhanced systems such as LODD. An example of this is Uniprot, which in addition to making all their data available as plain text or XML, also makes it available in the form of RDF triples. This makes it relatively trivial to link Uniprot records to other RDF-based data sources.

It is important to realize that, conceptually, semantically linking data sources is attractive as it allows a number of possibilities, primarily focusing on programs performing automated inferencing on linked data. But as of yet, most applications have focused on the use of RDF, OWL and other technologies as a way to first describe a domain (via domain ontologies) and second to link disparate data sources. These steps are no doubt necessary for further developments. What this means is that the use of these linked data sets to perform queries that will pull in related objects based on implicit relationships is a non-trivial task. In other words, specialized expertise is required to directly interact with these data sources and the associated technologies and therefore are probably not suited for use by bench scientists at this time.

5 The Tripod Platform for Integrated Browsing

As noted in Section 3.2, effectively addressing problems in chemical biology research requires the use of multiple databases. Manually dealing with disparate interfaces, merging common results and so on is tedious and error prone. As a result, integrated or federated solutions are much more useful for day-to-day work. The Tripod platform has been developed to address these issues by supporting the integration of multiple data sources in a flexible manner, and allowing users to query all the data in a simple and intuitive manner. The platform has already been deployed in one application, viz., the NPC Browser (Huang, Southall et al. 2011). This application is a browser for the NCGC Pharmaceutical Collection that allows users to examine small molecule structures as well as a multitude of meta-data from a variety of sources including literature (via PubMed), targets (via Uniprot), genes (via Entrez Gene), clinical trials (via ClinicalTrials.gov) and so on. A key feature of the Tripod platform (and exemplified in the NPC Browser) is the extensive linking between different data sources, allowing users to easily and rapidly identify items that maybe related to the initial query.

5.1 Integrated data sources

A key design feature of the Tripod platform is the ability to integrate a variety of data sources, in a piecemeal fashion. Thus, by default, the Tripod application comes with a few data sources bundled and linked to each other. Using the NPC Browser as an example, these would include a collection of small molecule structures, associated PubMed document id's

and abstracts, protein target information and so on. The key observation is that each data source is not necessarily stored in its entirety. However, there is no restriction on the size of the data source that can be included. Equally important is the ability to integrate local, possibly proprietary data sources into the application. Such data sources would appear in the sidebar (Figure 3) just like any other data source and would be included in queries seamlessly.

A core design feature of the Tripod platform is the interlinking of data sources at the record level. Thus, if it is known that a small molecule targets a particular protein, then there should be an explicit link between the records for the small molecule and the protein. The data sources bundled with the Tripod application primarily focus on common “objects” in chemical biology research: compounds, targets, genes, publications, pathways and so on. In addition, these data sources are pre-linked. The real utility arises when a user loads in their own dataset. Since the data sources for possibly related things (such as genes, protein targets etc.) are already available, the linking procedure can immediately identify the relevant records. Furthermore, since the default data sources themselves are already linked, it is feasible to identify “implicit” relationships, that is, relations between items that were not explicitly recorded - between user data and external data sources.

An important feature of the Tripod platform is automated updates. Thus, when corrections (from the original databases or via curation from users of the application) are made to the original data these updates must be made available to the users of the application. The Tripod platform has been designed to check for updates and when available automatically download and install them.

5.2 The role of curation

Errors are inevitable in databases and the key to ensuring integrity and enhancing the utility of a database is to provide mechanisms for rapid review and corrections. To support this, the Tripod framework allows updates to any field of a record. Currently, only chemical structure records are supported. The user can then submit the curated record to the central server. Locally, the database is updated and after review the curated record is updated in the central database and made available to all clients at the next database update. This mechanism allows rapid correction of chemical structure arrays, name to structure associations as well as filling in missing information. Other databases, notably ChemSpider also address the issue of curation. ChemSpider has a dedicated curation form and the results display provides an indication of whether a record has been curated or not. A similar approach has been taken in the Tripod framework (Figure 3).

5.3 Scaffold browsing

One of the use cases discussed in Section 3.2 was the use of scaffold-based methods for hit triage. More generally, a scaffold-based exploration of small molecule activities is an intuitive and efficient way of exploring large datasets. The Fragment Activity Profiler (<http://tripod.nih.gov/?p=206>) was developed to explore structure collections from the point of view of scaffolds, optionally annotated with activity information. The tool allows the user to explore activity profiles for a given scaffold across multiple targets or target families

(Figure 4). This tool is based on the ChEMBL database, but can easily incorporate user-generated data. The ChEMBL analysis is enabled by exhaustively enumerating scaffolds for all the molecules. Similarly, on importing user defined structures, fragmentation is automatically performed, allowing the analysis of private data in the context of external data sources (in this case ChEMBL). Thus given an assay in which a member of a scaffold has been observed to show significant activity, the Fragment Activity Profiler browser identifies compounds from ChEMBL that have the same scaffold and display their activities. Alternatively, the tool can characterize scaffolds in terms of their activities against different targets, and thus summarize the selectivity of compound series. This is especially easy with ChEMBL since structures (and therefore the derived scaffolds) are associated with targets via the assays they are tested in.

6 Summary

This article explores the challenges an experimental scientist is faced with when trying to access and prioritize information across the multitude of life science data sources. While there is a wealth of data available, keeping track of the available resources is problematic. While the primary, large databases such as those hosted by the NCBI and EBI are well known, many smaller, niche databases may only be known within a specific community. Given the interdisciplinary nature of chemical biology research, the information in such databases can be useful, yet locating the databases is a non-trivial task. Resources such as the Nucleic Acids Research database issue are quite useful in this respect.

At the same time, there is a lot of overlap between databases and it can be difficult to determine whether one database has missed some aspect of a query that is available in another database. While a number of approaches are available to address these problems, they all revolve around either federation of disparate data sources under a single interface or else integration of multiple sources (along with the concomitant job of merging identifiers and removing duplicates) into a single monolithic database. As a result, there are a number of data sources that represent such integration efforts. Some of these involve significant manual curation and are usually commercial products. However, there are an increasing number of free, public data sources such as DrugBank, PharmGKB and ChemSpider, that link a variety of information into a single interface and are partially (to varying extents) curated.

This paper has highlighted a number of use cases from high throughput screening and chemical biology that benefit from multiple data sources. In the absence of a pre-integrated solution, there is currently no easy way to manually integrate results from multiple data sources. The Tripod application supports the integration of data on chemical structures, genes, targets and publications but not in a fully automated manner. Currently, it provides a framework for integrating and linking multiple disparate data types, and as a result supports a browsing interface that lends itself to serendipitous discovery (Robas, O'Reilly et al. 2003; Ban 2006). In addition, the framework supports applications specific to chemical biology problems such as scaffold based browsing of structure collections.

While Tripod aims to hide much of the details of integration and search, other approaches are available that can help experimentalists with integrating data sources. One particular approach is the use of pipelining tools such as Pipeline Pilot, KNIME and Taverna. These tools allow the user to build up a pipeline of operations (e.g., “query Pubchem for this substructure”, “convert structures to InChI” etc.) using a graphical interface. While it simplifies many of the operations involved in integrating and mining disparate data sources by avoiding the need for low level programming, and provides much more flexibility than a pre-integrated solution, it still requires training on the part of the users and is likely not suitable for the casual user.

In summary, the modern experimentalist has huge amounts of information available to him or her, but distributed across a variety of data sources. While it is challenging to keep track of these databases and deal with the multiple data formats and user interfaces, a variety of resources have become available that simplify the querying and mining of multiple databases. Though the overview has covered a variety of data sources, covering various aspects of chemical biology, it is not meant to be a comprehensive review of all sources of information. A number of caveats regarding the use of multiple data sources have been described, with the hope that this enables researchers to assess the quality and validity of their results.

7 Online Resources

The list below is provided as a reference to the various online resources that have been discussed in this article.

- The Tripod website – <http://tripod.nih.gov>
- NPC Browser – <http://tripod.nih.gov/npc>
- Fragment Activity Profiler – <http://tripod.nih.gov/?p=206>
- NCGC chemical structure standardizer - <http://tripod.nih.gov/?p=61>
- Gene Literature Novelty Score – <http://assay.nih.gov/novelty.html>
- NCI Chemical Structure Lookup Service – <http://cactus.nci.nih.gov/chemical/structure>

References

- Apodaca R. Sixty-Four Free Chemistry Databases. 2011
- Ban TA. The role of serendipity in drug discovery. *Dialogues Clin Neurosci.* 2006; 8(3):344.
- Belleau, Fo; Nolin, M-A., et al. Bio2RDF: Towards a Mashup to Build Bioinformatics Knowledge Systems. *J Biomed Inform.* 2008; 41(5)
- Brennan RJ, Nikolskya T, et al. Network and pathway analysis of compound-protein interactions. *Methods Mol Biol.* 2009; 575
- Canny SA, Cruz Y, et al. PubChem Promiscuity: A web resource for gathering compound promiscuity data from PubChem. *Bioinformatics.* 2011 in press.
- Chen B, Dong X, et al. Chem2Bio2RDF: A Semantic Framework for Linking and Data Mining Chemogenomic and Systems Chemical Biology Data. *BMC Bioinformatics.* 2010; 11

- Dekker FJ, Koch MA, et al. Protein structure similarity clustering (PSSC) and natural product structure as inspiration sources for drug development and chemical genomics. *Curr Opin Chem Biol.* 2005; 9(3)
- Fourches D, Muratov E, et al. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J Chem Inf Model.* 2010; 50(7)
- Gao Z, Li H, et al. PDTD: A web-accessible protein database for drug target identification. *BMC Bioinformatics.* 2008; 9
- Halling-Brown MD, Bulusu KC, et al. canSAR: An Integrated Cancer Public Translational Research and Drug Discovery Resource. *Nucl Acids Res.* 2011 in press.
- Hishigaki H, Kuhara S. hERGAPDbase: A database documenting hERG channel inhibitory potentials and APD-prolongation activities of chemical compounds. *J Biol Databases and Curation.* 2011; 2011
- Huang R, Southall N, et al. The NCGC Pharmaceutical Collection: A Comprehensive Resource of Clinically Approved Drugs Enabling Repurposing and Chemical Genomics. *Sci Transl Med.* 2011; 3(80)
- Jessop DM, Adams SE, et al. OSCAR4: A Flexible Architecture for Chemical Text-Mining. *J Cheminf.* 2011; 3(1)
- Karopka T, Fluck J, et al. The Autoimmune Disease Database: A dynamically compiled literature-derived database. *BMC Bioinformatics.* 2006; 7
- Klein TE, Chang JT, et al. Integrating Genotype and Phenotype Information: An Overview of the PharmGKB Project. *Pharmacogenomics J.* 2001; 1
- Kramer C, Beck B, et al. A Composite Model for hERG Blockade. *ChemMedChem.* 2008; 3
- Kuhn M, Campillos M, et al. A Side Effect Resource to Capture Phenotypic Effects of Drugs. *Mol Sys Biol.* 2010; 6
- Kuhn M, Szklarczyk D, et al. STITCH 2: An Interaction Network Database for Small Molecules and Proteins. *Nucleic Acids Res.* 2010; 38:6. Database issue.
- Lachmann A, Xu H, et al. ChEA: Transcription Factor Regulation Inferred from Integrating Genome-Wide ChIP-X Experiments. *Bioinformatics.* 2010; 26(19)
- McNaught A. The IUPAC International Chemical Identifier: InChI. *Chemistry International.* 2006; 28(6)
- Metz JT, Hajduk PJ. Rational approaches to targeted polypharmacology: creating and navigating protein-ligand interaction networks. *Curr Opin Chem Biol.* 2010; 14(4)
- Murzin AG, Brenner SE, et al. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Bio.* 1995; 247(4)
- Ogata H, Goto S, et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl Acids Res.* 1999; 27(1)
- Oprea TI, Tropsha A, et al. Systems Chemical Biology. *Nat Chem Biol.* 2007; 3
- Osborne JD, Flatow J, et al. Annotating the human genome with Disease Ontology. *BMC Genomics.* 2009; 10(Suppl 1)
- Peri S, Navarro JD, et al. Development of Human Protein Reference Database as an Initial Platform for Approaching Systems Biology in Humans. *Genome Res.* 2003; 13(10):2371.
- Philippi S. Data and knowledge integration in the life sciences. *Brief Bioinform.* 2008; 9(6)
- Philippi S, Kohler J. Addressing the problems with life-science databases for traditional uses and systems biology. *Nat Rev Genetics.* 2006; 7(6)
- Posner BA, Xi H, et al. Enhanced HTS Hit Selection via a Local Hit Rate Analysis. *J Chem Inf Model.* 2009 ASAP.
- Reitz AB, Smith GR, et al. Hit Triage Using Efficiency Indices after Screening of Compound Libraries in Drug Discovery. *Curr Topics Med Chem.* 2009; 9(18)
- Robas N, O'Reilly M, et al. Maximizing serendipity: strategies for identifying ligands for orphan G-protein-coupled receptors. *Curr Opin Pharmacol.* 2003; 3(2)
- Samwald M, Jentzsch A, et al. Linked Open Drug Data for Pharmaceutical Research and Development. *J Cheminf.* 2011; 3(1)

- Schriml LM, Arze C, et al. Disease Ontology: A backbone for disease semantic integration. *Nucl Acids Res.* 2011 in press.
- Simmons K, Kinney J, et al. Comparative Study of Machine-Learning and Chemometric Tools for Analysis of in-vivo High-Throughput Screening Data. *J Chem Inf Model.* 2008; 48(8)
- Soh D, Dong D, et al. Consistency, Comprehensiveness, and Compatibility of Pathway Databases. *BMC Bioinformatics.* 2010; 11
- Steinbeck C, Kuhn S. NMRShiftDB -- Compound identification and structure elucidation support through a free community-build web database. *Phytochemistry.* 2004; 65(19)
- Su AI, Hogenesch JB. Power-Law-like Distributions in Biomedical Publications and Research Funding. *Genome Biol.* 2007; 8(4)
- Sun H. An accurate and interpretable bayesian classification model for prediction of hERG liability. *ChemMedChem.* 2006; 1(3)
- Tarcea VG, Weymouth T, et al. Michigan Molecular Interactions R2: From Interacting Proteins to Pathways. *Nucleic Acids Res.* 2009; 37:6. Database issue.
- Thorn CF, Klein TE, et al. PharmGKB: The Pharmacogenetics and Pharmacogenomics Knowledge Base. *Methods Mol Biol.* 2005; 311
- Wall DP, Pivovarov R, et al. Genotator: A disease-agnostic tool for genetic annotation of disease. *BMC Med Genomics.* 2010; 3
- Wang R, Fang X, et al. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J Med Chem.* 2004; 47(12)
- Warr W. Representation of Chemical Structures. *Interdisciplinary Rev Comp Mol Sci.* 2011; 1(4)
- Wishart DS, Knox C, et al. DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucl Acids Res.* 2008; 36
- Wishart DS, Knox C, et al. DrugBank: A Comprehensive Resource for In Silico Drug Discovery and Exploration. *Nucl Acids Res.* 2006; 34:672.
- Wu CH, Yeh L-SL, et al. The Protein Information Resource. *Nucl Acids Res.* 2003; 31(1)
- Yan SF, Asatryan H, et al. Novel Statistical Approach for Primary High-Throughput Screening Hit Selection. *J Chem Inf Model.* 2005; 45(6):1790.

Figure 1.
A count of new databases described in the yearly database issue of Nucleic Acids Research, over the last five years.

Figure 2.
A screenshot of the Entrez cross-database search interface highlighting the federation of multiple NCBI databases under a single query interface

Figure 3.
An annotated screenshot of the NPC Browser, an application based on the Tripod framework. The figure highlights the key features of the interface that contextualizes data by integrating and linking multiple datasources.

Figure 4. Screenshots of the Fragment Activity Profiler highlighting activity profiles for a single scaffold (left) and pairwise comparison of activity profiles for two scaffolds (right).

Table 1

A list of chemical databases, both public and commercial.

Database	Availability	SAR Information	Curated	Notes
GVK GoStar	Commercial	Yes	Yes	
Prous Integrity	Commercial	Yes	Yes	
PubChem	Free	Yes	No	
ChEMBL	Free	Yes	Yes	
PDSP	Free	Yes	Yes	Reports SAR data focusing on psychoactive compounds
STITCH2	Free	No	No	A collection of known and predicted interactions between small molecules and proteins
Drugable.com	Free	No	No	
ChEBI	Free	No	Partial	Its primary use is as a dictionary of small molecule
ChemSpider	Free	No	Partial	
canSAR	Free	Yes	Partial	Focused primarily on cancer drug discovery, integrates chemical screening data with RNAi, mRNA and 3D structural data
eMolecules	Free	No	No	Primarily a source for purchasing chemical structures
ZINC	Free	No	No	Contains purchasable compounds

Table 2

A list of pathway and interaction databases, that take into account small molecules. See Pathguide (<http://www.pathguide.org>) for a comprehensive listing of these databases.

Database	Availability	Curated	Notes
GeneGo	Commercial	Yes	
Ingenuity	Commercial	Yes	
Wikipathways	Free	Yes	
NCI PD	Free	Yes	
KEGG	Free	No	
STITCH2	Free	Partial	
HPRD	Free	Partial	
Reactome	Free	No	