## CHAPTER 1

# BRIDGING CHEMICAL AND BIOLOGICAL INFORMATION: PUBLIC KNOWLEDGE SPACES

PAUL A. THIESSEN
National Center for Biotechnology Information, National Library of Medicine,
National Institutes of Health, Bethesda, Maryland

WOLF-D. IHLENFELDT
Xemistry GmbH, Lahntal, Germany

EVAN E. BOLTON and STEPHEN H. BRYANT
National Center for Biotechnology Information, National Library of Medicine,
National Institutes of Health, Bethesda, Maryland

## 1 LANDSCAPE OF PUBLIC CHEMICAL (BIOACTIVITY) DATABASES BEFORE PUBCHEM

At the time of this writing, PubChem[1] is probably the most widely known publicly accessible chemical compound database on the World Wide Web (WWW, or just Web). It contains not only chemical structures, but also biological data linked to these structures. PubChem was launched in 2004, but it is certainly not the first freely available, Web-accessible database providing biological information on the Internet.

The biological data landscape is complicated by varying definitions of what classes of information should be considered as biological information. Do toxicity data constitute biological information? If yes, should a qualifying database contain actual measurements, or can this information be provided in distilled, abstracted formats, perhaps even as material safety data sheets (MSDSs) or simple handling classifiers? Do we simply

*Computational Approaches in Cheminformatics and Bioinformatics*, First Edition.
Edited by Rajarshi Guha and Andreas Bender.

consider biological information in the context of drug research, or is basic biological data (e.g., metabolic pathways) part of the picture?

The following descriptions of databases launched before PubChem should not be considered comprehensive but, rather, an editorially selected collection, highlighting novel features and the influence that these systems had on the development of later systems. Several sites have attempted to catalog all major Web-accessible chemistry databases (e.g., the Chembiogrid[2] resource), which the reader may want to consult for a broader picture. Additionally, an overview of chemistry and the Web in 1998 was published in a special issue of *Chimia*.[3]

The Protein Data Bank[4] (PDB), begun in the 1970s and available on the Web since the early 1990s, can be considered a grandfather of chemical structure databases, although with a rather peculiar and narrow focus. PDB stores and redistributes crystal structures of proteins and other biological macromolecules. This includes proteins with bound small molecules, information of high biological relevance. The actual structures have always been available for download, from basic FTP sites, shipped tapes or CDs, or the current Web interface. Nevertheless, small molecules and bioactivity data were never the principal focus of this database. Even today, the extraction of small ligand molecules from available data files remains a challenge, due to the particularly limited and often abused encoding standards employed. Only recently has PDB begun to provide nontextual ligand search capabilities. Link-outs to biological activities stored in external databases are still absent. PDB has stood the test of time and provides unique information but is rather isolated on the Web, despite numerous databases making the effort to establish relationships between PDB entries and their data (via unidirectional links).

Among the original small-structure chemistry databases making an entrance during the dawn of the Web, ChemFinder[5] by CambridgeSoft (development started in 1995) was probably the most influential and most professionally managed system. This was not, however, the first widely recognized small-molecule repository—that honor probably goes to the NIST WebBook[6] (online since 1996), but it contained only nonbiological data such as spectra and physical constants. ChemFinder pioneered many of the query and interface techniques still used today in Web chemistry databases, such as intelligent query parsing, structure search capabilities, and link-outs to secondary databases. Like PubChem (more details to follow), ChemFinder did not attempt to store all information located but, rather, linked to the original source. Because CambridgeSoft is the developer of the widely used chemical structure drawing program ChemDraw, ChemFinder was also designed as the showcase for the Web browser plug-in variant of ChemDraw. Using the ChemDraw plug-in, this was the first database to provide comfortable interactive drawing of structures for full-structure and substructure queries on the database, although at the expense of using a nonportable Microsoft Windows/Netscape-only interface (at the time of launch). Originally, ChemFinder was not specifically concerned about biological activity links. It indexed sites that the development team deemed important and indexible with the technologies available to the engineers, which included rather sophisticated chemistry-aware text-matching algorithms, allowing the establishment of database links even in the face of spelling variants and misspellings. The original ChemFinder database is no longer accessible. CambridgeSoft is relaunching it under the ChemBioFinder brand. The new release directly incorporates various drug databases, such as the *Merck Index*[7] and the National Cancer Institute (NCI) Developmental Therapeutics Program (DTP/NCI) cancer and antiviral screening data.[8]

The DTP/NCI database contents were prominent in the history of bringing biological data to the Web. This data set was first available on the Web via the NCI database browser[9] (currently in version 2). The first version of the NCI database browser was released in 1998, with about a quarter of a million structures from DTP/NCI. This compound set was collected over four decades but had only been accessible by an in-house system at NCI. The biological aspect of the database included the results of tumor cell line screenings of these compounds, measured on a collection of standard tumor cell lines. A smaller subset of compounds was also subjected to antiviral screens, with a special focus on anti-AIDS activity. The original compound data was (and is) problematic—many structures were registered without stereochemistry, and even the reconstruction of the connectivity of some structures is not always possible in an unambiguous way, due to the original coding of the in-house registration system.

The NCI database browser pioneered many important features. Among the Web structure databases of the time, it had the most sophisticated query system (even by today's standards), including features and abilities such as dynamically generated query forms (via JavaScript) and advanced tools to merge, manage, and store query and hit lists. Another important functionality in the design of this database was extensive export options for result sets with dynamic format conversion, enabling the use and reuse of the database contents for local projects. Until the advent of PubChem, this functionality was largely overlooked, with Web interfaces to public resources (even to this day!) designed with the single purpose of human browsing, with meager export capabilities—only parts of the records or a single full record at a time. Restrictive public resources with insufficient data filtering and export capabilities make the goal of reusing and reanalyzing public datasets very difficult to realize.

The NCI browser was among the first major chemistry database systems on the Web to implement a platform-neutral interface for structure searching and three-dimensional (3D) visualization. For structure input, it relied on the (then) newly released JME Java structure drawing applet,[10] an important development and a popular tool even today. Its result-display routines pioneered the use of dynamically generated GIF images of structures, where a query was displayed directly on the results using structure highlighting and other annotations depending on the query, a rarely found feature even now. For 3D visualization, the browser was first to support the export of structure models as virtual reality modeling language (VRML) files—at the time a highly promising general 3D display standard for the Web, but no longer well supported in the Web ecosphere. More common in chemical applications now are Java-based approaches such as JMol[11] for 3D chemical visualization. While the use of platform-independent approaches for public Web systems is now considered mainstream, at the time there was considerable dependence on external helper applications (e.g., RasMol[12]) and platform-specific plug-ins (e.g., MDL Chime[13]).

While the NCI database browser was a pioneer for the distribution of assay data, the Klotho database[14] (now defunct) was similar in that it was the first system to link biological pathway data with small molecules. Although not a direct successor, KEGG[15] (started in 2000) is now assuming its role. KEGG's PATHWAY database provides information about the role of small molecules in biological pathways, while the LIGAND database and its various sub-databases summarize data on chemical structures in the KEGG collection. A unique feature of KEGG is that it contains reaction information, linking the transformation of structures, although without an exact atom mapping (which the commercial database Biopath[16] has). Additional important databases in the biological pathway context are the Human Metabolome Database[17] (HMDB, online since 2004) and the BRaunschweig ENzyme DAtabase[18] (BRENDA, online since 2003).

PubChem is not the first public chemistry database supported by a long-term U.S. government sponsoring commitment. ChemIDplus,[19] which, like PubChem, is maintained under the umbrella of the National Library of Medicine (NLM), is older than PubChem. This database is important because it is considered one of the most extensive public toxicological information resources on the Internet. ChemIDplus contains nearly 400,000 records, many containing detailed toxicity information, and is linked to the NLM Toxicology Data Network (TOXNET) cluster of related databases, such as TOXLINE[20] (toxicology literature), GENETOX[21] (genetic toxicology), and CCRIS[22] (carcinogenicity and mutagenicity data).

A database that can be considered a direct precursor of PubChem is ChemBank,[23] launched in 2003 with support from the NCI Initiative for Chemical Genomics (ICG). ChemBank contains nearly 1.7 million chemical structures generated from vendor catalogs and filtered by various computational criteria relevant for drug design (e.g., rule-of-five[24] compliance, substructure and element exclusion, drug likeness score, and chemical diversity). These compounds are linked directly to biological screening results. To our knowledge, ChemBank is the first public chemical bioactivity database service supporting a dedicated application programming interface (API) for remote programmatic queries by applications other than Web browsers, through a set of Simple Object Access Protocol (SOAP) functions. This is similar to the more extensive interface that PubChem provides with its Power User Gateway[25] (PUG).

The landscape of publicly accessible databases with chemical and biological content has expanded since the launch of PubChem. Some databases involving sizable data curation efforts, such as ChEBI[26] (2005) and DrugBank[27] (2006), focus on drug and drug candidate information. Some databases involving sizable biological activity literature abstraction efforts, including BindingDB[28] (2001), PDB-Bind[29] (2004), KiBank[30] (2004), and BindingMOAD[31] (2005), focus on small-molecule binding constants. The earliest of these, BindingDB, is notable in that it allowed research groups to contribute data directly. The effective takeover by the European Bioinformatics institute (EBI) of a major commercial bioactivity knowledge base—the BioFocus DPI StARLITe[32] database—is a recent development that may be considered nothing short of remarkable given the breadth and depth of bioactivity information directly relevant to drug discovery. The StARLITe database, integrated into the public knowledge bases at EBI (e.g., ChEBI), will be a welcome addition to the publicly accessible space, with more than 2 million bioactivity data points abstracted from 12 journals for about 1500 drugs, 10,000 drug candidates, and 450,000 drug leads. This may be a sign of things to come as public knowledge bases grow in size and quality, potentially limiting the space in which commercial vendors can operate.

## 2 PUBLIC DATABASE INTEGRATION EFFORTS

Unlike the databases mentioned above, PubChem is neither an originator of bioactivity information (e.g., DTP/NCI or ChemBank) nor is it a curation or literature abstraction effort (e.g., ChEBI, KEGG, or PDB-Bind). It has no tiered data access scheme, no log-in requirement, and no restriction on who may contribute. PubChem is an open repository, depending entirely on external contributors for its content. PubChem was originally funded as a part of the Molecular Libraries Program (MLP), a component of the National institutes of Health (NIH) Roadmap.[33] This program includes the Molecular Libraries Probe Production Center Network (MLPCN), consisting of grant-supported experimental laboratories, and a shared compound repository referred to as the Molecular Libraries Small Molecular Repository (MLSMR), offering biomedical researchers access to chemical samples. The MLPCN is the successor to the Molecular Libraries Screening Center Network (MLSCN) from the initial MLP pilot phase.

PubChem archives the molecular structure, bioassay data, and annotations from the MLP and third-party depositors. PubChem provides search, retrieval, and data analysis tools to optimize the utility of information collected. Also, PubChem imports other public sources of chemical structure and bioactivity information, and integrates this with data contributed and NIH biomedical knowledge bases (e.g., PubMed,[34] MMDB,[35] GenBank,[36] MeSH,[37] DailyMed[38]). The primary aim of PubChem is to provide a public online resource of comprehensive information on the biological activities of small molecules, accessible to molecular biologists and to computational and medicinal chemists.

ChemBank, perhaps the public data system closest in content and form to PubChem, is similar in various ways. Both databases contain large amounts of small-molecule structures and associated bioassay data. Both databases provide tools to search and analyze these data. Both databases have similar stated goals of providing freely available information. There are considerable contrasts, as well. ChemBank is not an open repository. ChemBank data are generated locally at the Broad Institute, giving ChemBank complete control over the content and verbosity (i.e., the “rawness”) of biological screening data provided. PubChem takes chemical structure, screening information, and other data from many organizations, including ChemBank, each with its own ideas on what is necessary to communicate experimental results to the public. ChemBank embargoes all new data from public access as a matter of policy. As such, ChemBank can be accessed either in public form or via the data-sharing agreement (DSA), which grants access to both public and embargoed data. PubChem normally releases data immediately once the depositor is satisfied with the data import accuracy and overall presentation within PubChem (using the PubChem Deposition Gateway[39]). It is possible for a PubChem depositor to put data on hold: for example, to synchronize the release of PubChem records with the publication of a paper or announcement of a new resource, but PubChem does not hold data as a matter of policy and there is no way for any user (even the originating depositor) to search or analyze such data, as with the ChemBank DSA. Considering that PubChem is run by a government agency, it is restricted from tracking public data users so as not to violate privacy laws, precluding the use of any tiered or collaborative data access model requiring a login, such as ChemBank DSA.

There are strengths and weaknesses in both these database models. PubChem's open model prevents its use for selective release prior to publication, whereas selective release is often the preferred method in the highly competitive scientific "publish or perish," intellectual property–centric environment that exists in both the private and academic sectors; but PubChem lets anyone integrate and cross-link their own data with those of many other data originators and biomedical knowledge bases with minimal effort, as most of this analysis is performed automatically upon deposition. The ChemBank collaborative model provides freedom to decide who gets to see what and when, but it is restricted to data originated by ChemBank collaborators. Although many other parallels and contrasts can be made, it is really the data collection policy that sets PubChem apart from ChemBank and other databases. Given its unique nature as a freely available, public, and open archive, future discussion will focus primarily on PubChem: its contents, methods of integrating data from disparate sources, and the caveats involved in such a system.

## 3 DESCRIPTION OF DATA CONTENTS

There are two primary aspects of a chemical database: the policy and procedures by which the database is populated, and the actual data it contains. Some are curated manually; that is, records are entered and checked by a human (e.g., KEGG or BioCyc). On the opposite end of this spectrum is PubChem, whose data entry is entirely automated, with data provided by depositors treated as is. ChemSpider[40] is between the two, where much data acquisition is automated, but individual records are open to manual adjustment. One may legitimately argue the value of the various approaches, but the biggest factor here is a combination of the size of the database and the human-hours available for manual data entry and validation, through either direct staffing or a wiki-style open system.

PubChem collects information from depositors via the PubChem Deposition Gateway largely to the degree of detail they are willing to provide. For substances, the only field required is an external registry ID; however, a rich set of information may be and often is provided, including a chemical structure in Structure-Data File[41] (SDF), the Simplified Molecular Input Line Entry System[42] (SMILES), or IUPAC International Chemical Identifier[43] (InChI) format; uniform resource locators (URLs) to external Web-enabled resources; substance synonyms; comments; and annotations in the form of cross references to NCBI knowledge bases (including PubMed, GenBank, MMDB, Gene, Taxonomy, OMIM, and Probe). Many of the databases mentioned earlier are also contributors to PubChem, including BindingDB, ChEBI, ChemBank, ChemIDplus, DTP/NCI, KEGG, and MMDB. Each PubChem depositor may classify the types of information they provide (e.g., biological properties, substance vendor, chemical reactions). Although the contribution to PubChem by individual depositors varies, a deposition often serves to inform PubChem users that additional information is available at the depositor's website.

Visited on average by more than 70,000 unique users per day, PubChem contains (at the time of writing) over 85 million substance descriptions from over 160 depositors, 500,000 bioassays with more than 140 million substance outcomes originating from over 45 depositors, and more than 30 million unique chemical compounds. A current list of PubChem contributors and record counts is available.[44] The data content is still growing, especially with respect to assays and assay test outcomes; the number of unique chemical compounds grows more slowly as PubChem's coverage of known chemicals increases.

PubChem is organized as three distinct databases: substance, compound, and bioassay. The substance database contains depositor sample descriptions, necessarily including any chemicals tested in any assays by that depositor. Compound records are created from PubChem substances through the means of standardization or normalization of deposited chemical structure information. (Standardization is described in more detail later.) As such, the PubChem compound database is derived from the PubChem substance database and represents the overall unique chemical structure content of PubChem. Each PubChem compound record consists of a fully defined chemical structure (no implicitly defined valence, hydrogens, or aromatic bond definitions), computed properties (e.g., molecular weight, molecular formula, hydrogen-bond donor counts), and identifiers (e.g., IUPAC name, SMILES, InChI).

PubChem bioassay records are more complex than substance definitions. They consist of two logical parts, a definition and associated data. Unlike PubChem substance records that have minimal requirements, bioassays require a more complete record from a depositor, with separate sections for description, protocol, target definition, comments, readout definitions, URLs, and annotations. After a bioassay definition is created by a depositor, data for the assay may be given for each substance tested in the bioassay, including an activity outcome (for partitioning purposes, e.g., active, inactive, inconclusive), activity score (for sorting purposes, e.g., a value between 0 and 100, with larger values indicating more active or more interesting results), URL, and annotations. A PubChem bioassay record may be configured to provide “panel-style” results, such as in the case of phenotypic assays or selectivity profiling assays where there are multiple outcomes. These panel assays may have multiple sets of targets, outcomes, and results defined, providing for a compact and data-rich representation.

Bioassays may contain a target definition, typically a specific protein. These allow (indirect) cross-links to be formed between GenBank entries and PubChem substances and compounds. Thus one may group proteins of interest (e.g., with sequence analysis tools such as BLAST[45]) and thereby discover sequence families whose members have been targets of PubChem assays and chemicals that have been found to interact with these proteins. Assay targets may also be nucleotides and may have cross-links to genes and taxonomy. PubChem precomputes similarity between known assay targets, allowing one to cluster or navigate by target similarity. One may also search PubChem assay targets by sequence, using a specialized subset of the online NCBI BLAST resource.[46]

Another important source of biological information for PubChem records is the NLM Medical Subject Heading (MeSH) classification. This ontology of controlled vocabulary is used to index articles in the biomedical literature, containing many chemical names and their known biological function. For example, the term *aspirin* has a description, a list of synonyms (acetylsalicylic acid, acylpyrin, etc.), pharmacological actions (anti-inflammatory agents, nonsteroidal, fibrinolytic agents, etc.), a place in a tree of chemical structure classification (salicylic acids), and categorized links to PubMed journal articles that refer to aspirin. PubChem automatically matches the names supplied by depositors to the MeSH terms and links to MeSH and PubMed as appropriate. The reciprocal links are also present; that is, through a simple link in Entrez, one may easily get a list of all PubChem substances or compounds that share the MeSH pharmacological action “fibrinolytic agents,” which includes the record for aspirin. Similarly, one may find PubMed articles that refer to a particular chemical, or conversely, list chemicals referred to in an article. PubChem thus uses MeSH as both an annotation and classification system, and as a direct link between chemicals and the biological literature. Although not perfect—not all chemicals mentioned in an article are indexed, and not all articles in PubMed are MeSH-indexed—integration of PubChem and MeSH provides a valuable tool to learn about the biological function of small molecules and the literature available.

As mentioned above, the Protein Data Bank (PDB) is a direct source of detail on small molecules’ three-dimensional interaction with proteins.

NCBI maintains an effective mirror of PDB called the molecular modeling database (MMDB), from which PubChem extracts the small molecules (ligands). This process is fully automatic, and some details and problems with that are discussed more fully below. MMDB ligands provide a direct link between PubChem records and Entrez's structure database, and thence to proteins and nucleotides in GenBank.

Links to a wide variety of other types of biological data are indirectly available through PubChem in records imported from other public databases. For example, one may find information on drugs and metabolic pathways (e.g., KEGG and BioCyc[47]), toxicology (e.g., ChemID Plus, EPA/DSSTox[48]), cancer screening (e.g., DTP/NCI[49]), anti-HIV screening (e.g., DTP/NCI, NIAID[50]), imaging agents (e.g., MICAD[51]), scientific journals (e.g., *Nature Chemical Biology*[52]), NMR data (e.g., NMRShiftDB[53]), and protein-binding affinity (e.g., BindingDB), just to name a few.

## 4 TECHNICAL ASPECTS OF BIOACTIVITY KNOWLEDGE BASE DATA INTEGRATION

In many ways, collection of information from data contributors is the easy part. There are many aspects to integration of such data to maximize navigation and interpretation. Although not meant to be an exhaustive discussion on the topic, describing concepts and ways in which this is achieved within the scope of PubChem is the primary goal of this section.

## 5 CHEMICAL STRUCTURE HANDLING

Chemical structure is often the key to joining disparate data sources. Each depositor can have completely different textual names which they use to refer to a substance but still have the same chemical structure. A simple text search of these two records would not reveal that they are related. If the chemical structure is the same, the list of textual names may be combined along with any other information known about the two substances. This is the power of the chemical structure.

A primary step toward using chemical structure in an integrated biochemical resource is determining when two or more records are actually referring to the same chemical, so that the records may be linked together with that chemical as the commonality. To do this correctly requires knowing in full detail the exact chemical structure being represented. This can be problematic when integrating multiple heterogeneous sources, all of which may have their own means of describing small molecules, in a greater or lesser level of detail. In this section we present some of the difficulties in comparing chemical species provided by disparate depositors and the concepts behind the tools that PubChem uses to provide effective integration and cross-linking of chemical structure information.

## Standardization of Chemical Structure Representation

There are no universally adopted rules on how to convey chemical structure data. As such, each organization (or chemist within an organization) is free to adopt chemical drawing conventions arbitrarily. PubChem receives data from many different organizations, compounding the issue, as such drawing conventions may conflict between depositors. Specialized processing is required to normalize the representation of chemicals in a consistent way to allow data from different depositors to be combined and integrated when the two records are associated with the same chemical structure. Such standardization processing requires modification of the original data provided by the depositor.

Making such changes directly to the chemical structures at the time of import into PubChem would violate the very archival nature of PubChem, intent on preserving data provided by depositors. Furthermore, it may be rather unsettling to the contributing organization to have their data modified, especially if there is a mistake in such processing. To balance the competing demands of preserving original information and the need to unify chemical structure representation, to the best of its ability, PubChem retains the deposited chemical structure information and, after successful normalization processing, associates one or more PubChem compound records with the substance record. This allows PubChem processing rules to change as a function of time while preserving the integrity of the original deposited data.

Actual chemical structure processing steps used by PubChem involve a series of verification steps, including atomic element checks, functional group normalization, and atom valence checks; standardization steps, including valence-bond representation normalization (for tautomer-form invariance), aromaticity normalization (for VB-form invariance), $sp^2/sp^3$ stereochemistry detection (including systems involving allenes, allene ring equivalents, or free electron pairs), and explicit hydrogen assignment; and mixture component processing steps, including covalent unit detection, proton-based neutralization (when applicable), and parent assignment. Components detected during processing are individually reprocessed in the same way. The final processed structure and unique components are associated with the substance record as standardized compounds. If a substance fails normalization processing, no PubChem Compound record is associated with the substance. PubChem provides a standardization service[54] to allow users to standardize chemical structures: for example, to determine what PubChem Compound record(s), if any, are associated with these chemical structures, an especially important step when seeking to integrate private data with PubChem.

## Identity Groups

One of the compelling reasons for creating a chemical biology resource consisting of various depositors is the ability to compare bioactivity data between depositors. Standardization of chemical structures into a uniform representation is a big first step in this direction, but it does not resolve the issue completely. Different depositors may provide variable levels of chemical structure detail of what many would consider to be essentially the same structure, depending on purpose. For example, two depositors may have the same chemical structure as far as connectivity is concerned, but one depositor may have provided a complete stereochemical description whereas the other did not. An additional concept employed by PubChem to locate related structures is "identity groups" to provide variable degrees of "sameness" at the level of connectivity, where chemical structure connectivity (atoms/bonds) is identical but variation may occur at the isotopic and stereochemical levels; stereochemistry, where chemical structure connectivity and stereochemical centers are identical but different isotopic forms are allowed; isotopes, where chemical structure connectivity and isotopic form are identical but variation is allowed in stereochemical centers; exact, where chemical structures must be identical at all levels (connectivity, stereochemical, and isotopic); and any tautomer, where a more liberal tautomeric identity representation is used that considers tautomerism that occurs under mild (acidic, basic, or temperature) conditions, allowing for variability at the connectivity, stereochemical, and isotopic levels. (Internally, these groups are implemented as constant-width 64-bit structure hash codes,[55] which are compared easily and efficiently and are faster to compute than canonic linear structure representations.)

## Mixtures and Parents

Chemical structures in PubChem are often provided in various forms, salts with different counterions, formally charged, neutralized, and so on. PubChem standardization detects mixture components and associates these components with the record for the entire mixture. For example, the chemical structure of sodium acetate will be linked to the components sodium and acetic acid. Thus, one may “expand” or “collapse” a list of chemical structures to explore mixtures or components of compounds in which one is interested. When considering biological activity, it often does not matter in what form a compound is tested (salt or nonsalt), such that it may be useful for bioactivity comparison purposes to consider acetic acid and sodium acetate as being the same structure. To handle this effectively, PubChem uses the concept of a parent compound. In the case of sodium acetate and acetic acid, acetic acid is assigned as the parent. Not every compound will have a parent. Complex mixtures containing similarly sized organic moieties, or purely inorganic mixtures, do not have parents because here the concept of “parent” has ambiguous meaning. The notion of the parent compound allows a compound search to be expanded to a larger number of chemical structures that are presumably biologically equivalent; simultaneously, bioactivity data may be collapsed to a smaller number of equivalent compounds.

### Similar Compounds

The combination of the paradigms of parent compound, identity group, and standardization provide very powerful means to navigate bioactivity information of “same” chemical structures. PubChem provides similarity neighboring relationships for structures that are not the same but are very similar: for example, analogs in a Markush-type series. This allows a user to expand the scope of compounds or substances considered to include additional chemical records that are structurally similar and thus that may have similar biological activity. These neighboring relationships, precomputed for each compound record, are equivalent to a PubChem 2D similarity search using the Structure Search tool[56] at a 90% threshold.

## 6 STRUCTURE SEARCHING

PubChem’s compound database is searchable by all standard 2D structure search methods: full structure (internally performed via hash code comparison), substructure and superstructure, structural similarity (using the classical approach with screening bit vectors for acceleration and similarity score computation), and chemical formulas. The similarity comparison uses a special boosting scheme to assign scores above 100% for identities not normally distinguished by this algorithm: 104% for full isotope and stereo identity, 103% for either stereo or isotope match but not both, 102% for connectivity identity, and 101% for tautomer identity.

## 7 TEXT SEARCHING

In this section we outline some of the problems associated with plain text searching in a chemical structure database. There is a notorious variation in how molecules are named, even among knowledgeable chemists; and, of course, there is always the possibility of error. All of these problems are multiplied in a large chemical database with redundant structures, and even more so in a database such as PubChem, which brings in data from numerous sources. Here we describe some weighting strategies used by PubChem to increase the likelihood that a text name search will result in the correct structure.

### Search Fields and Errors

An immediate problem is one of interface. If there is a single text box for user data entry, the vast majority of users will simply enter the name there without understanding the underlying search details. For example, PubChem's entry for acetaminophen (Tylenol, CID 1983) has the name "aspirin-free Anacin" as one of the synonyms supplied. Hence, a completely unrestricted text search in PubChem for the term "aspirin" will bring up the record for acetaminophen. It is possible to avoid this to some degree by narrowing the search appropriately using the "CompleteSynonym" index; a search for "Aspirin[CompleteSynonym]" in PubChem will be more reasonable but still finds cases where the record is a mixture of multiple chemicals. Take, for example, CID 24847967, where the depositor has misused the synonym field when submitting to PubChem and has specified the names of the chemical components of the mixture—aspirin and oxycodone hydrochloride—as separate names for the mixture itself.

As the complexity of the molecule increases, so does the chance that different sources will disagree on—or simply make a mistake about—the details of the structure. Stereochemistry is a common example of this sort of problem; searching "vancomycin" in PubChem's Compound database (again without restriction) currently yields 71 structures; if narrowed to synonyms that are exactly "vancomycin," the search still yields six structures. The correct structure (CID 14969, according to ChemSpider) has 28 neighbors with the same atoms and bonds but different stereochemistry, four of which are called "vancomycin." Even for a human chemist, determining what the correct structure is may be problematic, and for a computer alone, nearly impossible.

### Weighting

With all of these problems, one might ask: How can a PubChem search for “aspirin” show the correct structure of aspirin as the first result? (Keeping in mind that unlike many other databases, PubChem has no mechanism for a curator or expert to step in and identify one manually.) The answer is that PubChem attempts to weight particular terms more heavily than others, and sorts the records based on the weight of the terms that are matched to the original query. This is not a novel strategy, but the important part is how the weights are determined.

Here PubChem can actually take advantage of its diversity of information sources using a voting scheme to weight common names more heavily. That is, if many depositors supply the name “aspirin” for a given chemical structure but only a few supply the name “acetosal” for that same structure, the name “aspirin” gets a higher weight and a search for “aspirin” brings to the top the record for which the most depositors gave that name. PubChem gives a name only one “vote” per source, not one vote per source per record; this prevents sources that provide many records of the same structure from falsely imbuing too high a weight, such as MMDB, which contains hundreds of heme ligands called “Hem”—not a highly informative name.

PubChem also weights a name based on how many upper- and lowercase letters, numbers, and symbols (dashes, primes, etc.) appear in the name. The exact formula is not terribly important and is subject to change and so is not given here, but the effect is that more readable names, such as “aspirin,” get higher weights, while short, long, or numeric names get lower weights (e.g., “Sine-Off Sinus Medicine Tablets-Aspirin Formula”).

## 8 IMPORTING DATA FROM NONCHEMISTRY SOURCES

Some public scientific knowledge bases (e.g., PubMed) are devoid of well-defined chemical structure content (i.e., atom or bond description), containing only references to chemical names. In some cases, these resources have a focus other than chemistry (e.g., biological macromolecules, pathways, diseases). To bridge this gap between these knowledge bases, when detailed chemical structure information is simply unavailable, solutions must be created. Chemical name matching is an obvious choice: simply looking up the chemical name and, if there is an exact match, assigning the chemical structure to that data record; however, this problem is far from solved and has many caveats.

## Deriving Chemical Structure from Chemical Names

Deriving a chemical structure from a chemical name is nontrivial. A single chemical structure can have many chemical names. For example, in PubChem, aspirin (CID 2244) has more than 300 synonyms, being a mix of IUPAC or CAS names (of different generations), common names (some in different languages), product names, various registry names, and so on. A chemical name can also match multiple records. Again, using PubChem, an exact search for "aspirin" currently returns 15 results; one result is aspirin, another is an isotopic-labeled form of aspirin, and the rest are mixtures where aspirin is a component. Although a match may be found to a chemical name, that match may not be desired. One of the 15 exact matches to "aspirin" in the example above, mentioned previously, consists of a mixture of aspirin and oxycodone. It is also possible that the particular chemical name is not found even though the structure represented by the synonym may be available. These caveats attempt to stress how simple text matching of chemical names may readily provide false positives and false negatives that could be difficult to address in an automated fashion.

One may attempt to construct some controlled vocabulary for name matching, where a list of authoritative names are linked to a known structure. This is the method used by NCBI's MeSH database, where expert analysts match references to chemicals in the biomedical literature to particular entries in the MeSH database. CAS registry numbers[57]are also used for this purpose, although this is not as straightforward as it appears to be, because frequently, multiple CAS numbers have been assigned: to stereoisomers, isotope label variants, mixtures, formulations, extracts, and so on.

To provide a bridge between the text and chemical structure worlds when depositions are text only, PubChem provides the ability to generate chemical structures from chemical names at deposition time in three primary ways: when a synonym is a PubChem Compound identifier (e.g., CID2244); via MeSH, when a synonym matches one found in a MeSH record and when PubChem assigns just a single PubChem Compound record to the MeSH record; and using name-to-structure software (LexiChem[58]). Chemical structures generated in this fashion are not considered to be a part of the deposit record, as they may be updated and are annotated to the user as being derived. An astute reader will notice that simple chemical name lookup is not one of these methods; this is in part due to the aforementioned caveats.

The method using a PubChem identifier as a synonym is the most reliable, as the depositor already performed the curation step of matching their record to one existing within PubChem. The MeSH names are a controlled vocabulary used for name matching where expert analysts match references to chemicals in the biomedical literature to particular entries in the MeSH database. Provided that the link between CID and MeSH records is accurate, and if the depositor provided the correct synonym, this method is likely to provide good results for common names of substances. Name-to-structure software, although straightforward to use, has its own caveats. IUPAC (or IUPAC-ish) names have different styles and may contain some ambiguity, such that even different software packages or software versions may give different names for the same structure and different structures for the same name. Factor in the possibility of error when such names are generated by hand and IUPAC names begin to seem less reliable then some assume.

### Incomplete Structure

The Protein Data Bank (PDB) is a classic case of a biological database resource with incomplete detail in small-molecule structure. Most PDB structures do not contain hydrogen, bond orders, or formal charges, as these are simply not part of the format. These details need to be inferred from the 3D geometry, using bond lengths, angles, and torsions to arrive at the atomic hybridization. This is accomplished automatically in PubChem with the help of the OEChem toolkit.[59] It is not perfect, however, because of ambiguities in the data such as 1UA0's "AF" ligand (SID 26711741), which comes out of OEChem (version 1.5.1) with a pentavalent carbon or when attempting to distinguish between the $NADP^+$/NADPH or FAD/FADH2 redox pairs, which have planar ring systems differing only in charge and the number of hydrogens, both of which are unspecified in the PDB format.

### Literature extraction

To get chemical information from biological or other existing sources that have not maintained an explicit associated database of chemicals, it is often necessary to extract this information after the fact. Recognizing and cross-linking chemical references in biological journals and patent literature is an important and active field of development. The usual dichotomy exists between automated and manual extraction of this information—meaning, is the text being processed by a computer or a human? Even determining which words in the text refer to chemicals is not a trivial problem, let alone ensuring that the right chemical structure is matched in each case. Especially with patents, both the results and the technology used may be proprietary. Published materials may include chemical structure drawings, which somehow need to be entered into a computer database. Technology commonly called chemical OCR is being developed on a number of fronts[60,61] to enable a computer to convert drawings automatically into detailed chemical structure representations.

## 9 CASE STUDY AND LESSONS LEARNED

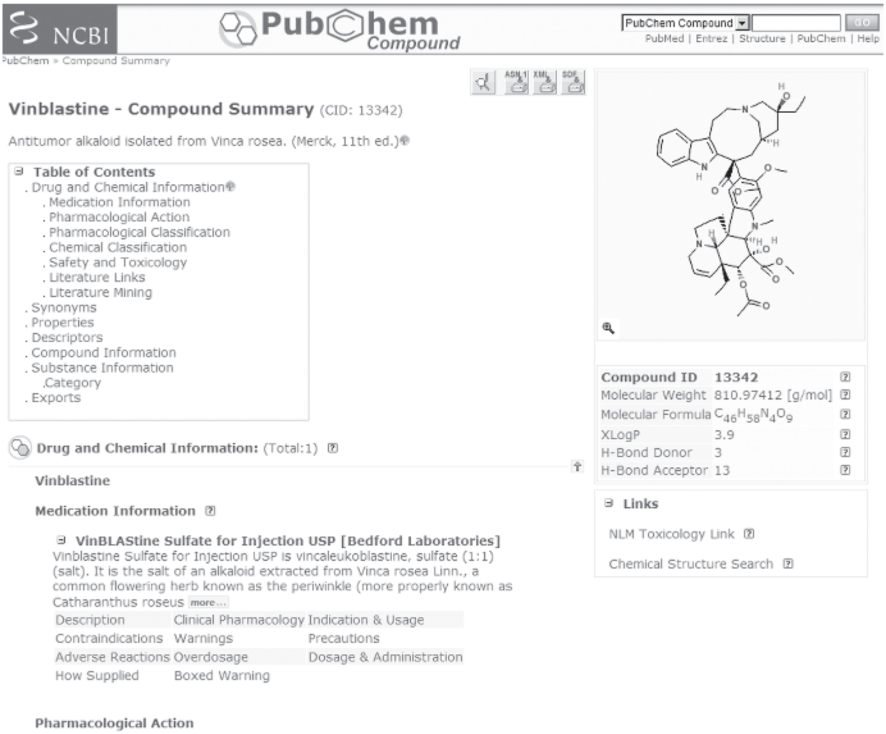
We have presented a variety of technical challenges to building a chemical database and have used PubChem as an example of how to approach some of these problems. But in any database there will be errors, and one key to using a database properly is discerning whether errors originate in the data itself or in the database's particular infrastructure and algorithms. Let us keep these points in mind when considering the following example: how one might attempt to extract information on structure and biological activity from PubChem, based only on a chemical name.

When reading a journal article where detailed chemical structure is not present, we come across the name "vinblastine" and would like to discover what is known about this chemical. There are really two basic questions to explore here: What is the "correct" chemical structure of vinblastine, and what do we know about its biological activity? Are there relevant bioassay results in PubChem?

A simple unrestricted text search in the PubChem Compound database for this word will result in a list of 25 different compounds (at the time of this writing), all varying slightly in stereochemistry, salt form, or even basic formula and connectivity. As mentioned before, PubChem attempts to prioritize the search results so that structure most likely to be correct comes first in the search result. This is not a perfect algorithm, however, and is subject to the overall accuracy of the records from the numerous depositors (data sources) who have provided PubChem with its information. One might compare to, say, ChemSpider, in which a search for "vinblastine" results in a list of five structures. Careful comparison of these structures shows that the first structure from PubChem (CID 13342) exactly matches the first structure from ChemSpider (ID 12773), but ChemSpider does not seem to claim that the first record is the correct structure—it may just be coincidence.

Looking more closely at the variety of PubChem results, one sees in particular the structure of vindesine, also called (by MeSH) a "vinblastine derivative" or analog. Vindesine records are thus found by an unrestricted search for vinblastine, as PubChem makes MeSH descriptions part of the search. This is why it is important when doing a very specific search to narrow the search to appropriate fields. Using PubChem's CompleteSynonym index, only records where the search term exactly matches the (entire) name in a record will be found. So, searching PubChem Compound for "Vinblastine[CompleteSynonym]" results in 5 records, and excludes structures such as vindesine and vinblastine sulfate. Still, it seems unlikely that there are that many variants of this structure, as natural products such as this one tend to occur in only a single stereochemical form. These searches alone do not narrow the list enough to conclude which is the correct structure.

Putting aside (for the moment) the question of structure identity, let us turn to biological activity. As depicted in Figure 1, the compound at the top of the list in the initial search above (CID 13342) has a variety of information linked to it underneath the "Drug and Chemical Information" heading, including MeSH classification, DailyMed drug information, and safety and toxicology links. However, this particular record is not linked to any bioassays. One could go back to the exact name search and examine those records. But this might miss structures tested in assays but for which,


NCBI
PubChem
Compound
PubChem Compound
GO
PubMed | Entrez | Structure | PubChem | Help
PubChem » Compound Summary
ASN.1
XML
SDF
H
O
N
H
N
H
O
O
O
N
H
N
H
H
O
O
O
O
O

**Vinblastine - Compound Summary** (CID: 13342)

Antitumor alkaloid isolated from Vinca rosea. (Merck, 11th ed.)

**Table of Contents**
. Drug and Chemical Information
  . Medication Information
  . Pharmacological Action
  . Pharmacological Classification
  . Chemical Classification
  . Safety and Toxicology
  . Literature Links
  . Literature Mining
. Synonyms
. Properties
. Descriptors
. Compound Information
. Substance Information
  ,Category
. Exports

**Drug and Chemical Information:** (Total:1)

**Vinblastine**

**Medication Information**

**VinBLAStine Sulfate for Injection USP [Bedford Laboratories]**
Vinblastine Sulfate for Injection USP is vincaleukoblastine, sulfate (1:1) (salt). It is the salt of an alkaloid extracted from Vinca rosea Linn., a common flowering herb known as the periwinkle (more properly known as Catharanthus roseus more...

| | | |
|---|---|---|
| Description | Clinical Pharmacology | Indication & Usage |
| Contraindications | Warnings | Precautions |
| Adverse Reactions | Overdosage | Dosage & Administration |
| How Supplied | Boxed Warning | |

**Pharmacological Action**

| | |
|---|---|
| **Compound ID** | **13342** |
| Molecular Weight | 810.97412 [g/mol] |
| Molecular Formula | $C_{46}H_{58}N_4O_9$ |
| XLogP | 3.9 |
| H-Bond Donor | 3 |
| H-Bond Acceptor | 13 |

**Links**

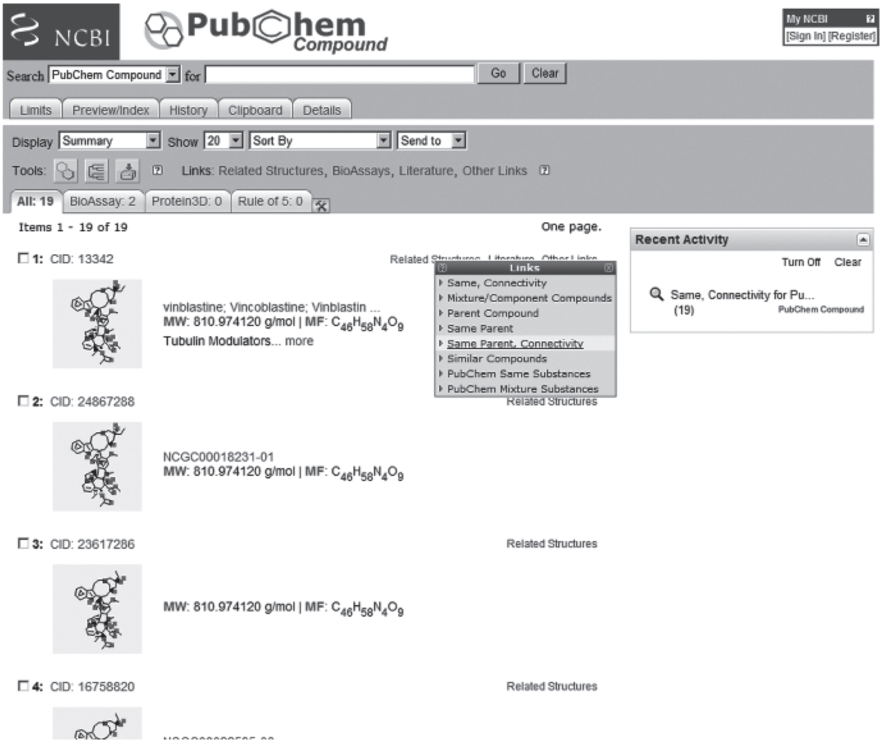
NLM Toxicology Link

Chemical Structure Search

**FIGURE 1** Partial view of the PubChem summary page for CID 13342, the nonsalt form of vinblastine.

for whatever reason, the name "vinblastine" was not given. To find such structures, one may use the PubChem links to compounds of the same chemical connectivity, found under the "Compound Information" heading on the summary page. From CID 13342, this link leads to a list of 19 compounds, 15 of which are called vinblastine, and two of these are tested in assays. However, since other ionic forms of the same structure are likely to have the same biological activity, one might better use the "same parent" links to expand the search; from an Entrez summary list from an earlier search, following the "Same Parent, Connectivity" link from the "Related Structures" pop-up menu for CID 13342 (see Figure 2), one arrives at a list of 46 compounds, seven of which have bioassay results. Indeed, at least one of these compounds with assay information (CID 16757894) does not have the name "vinblastine" associated with it at all.

From this Entrez result of 46 structures, let us use the bioactivity analysis tool to examine the bioassay results in more detail. Say that we are interested in compounds that are active in some assays. Activating the BioActivity Analysis button (containing two hexagons) near the top left



Items 1 - 19 of 19

□ 1: CID: 13342

vinblastine; Vincoblastine; Vinblastin ...
MW: 810.974120 g/mol | MF: $C_{46}H_{58}N_4O_9$
Tubulin Modulators... more

□ 2: CID: 24867288

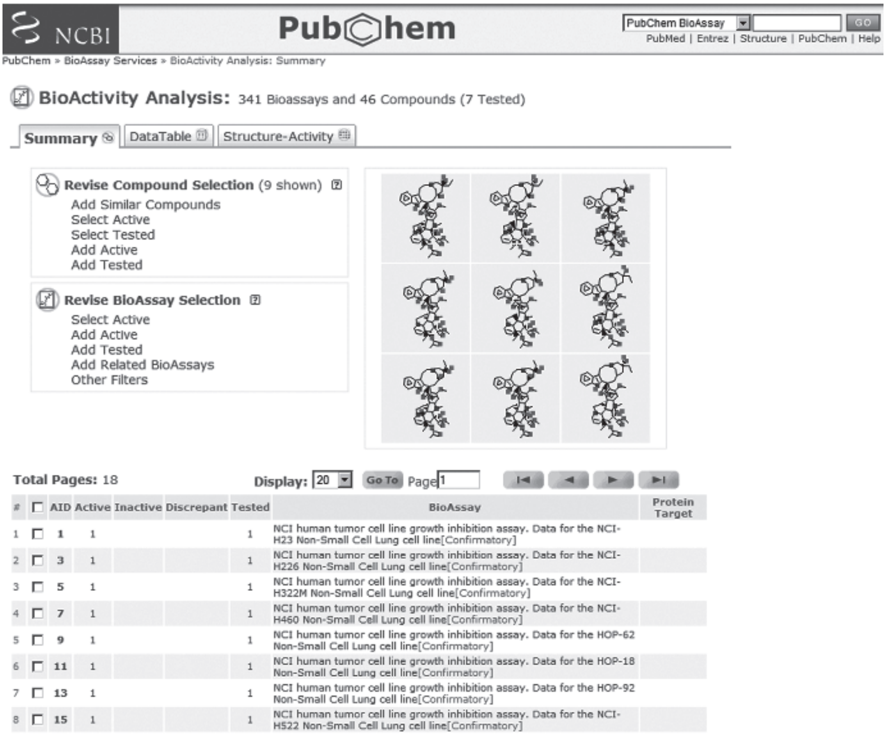
NCGC00018231-01
MW: 810.974120 g/mol | MF: $C_{46}H_{58}N_4O_9$

□ 3: CID: 23617286

MW: 810.974120 g/mol | MF: $C_{46}H_{58}N_4O_9$

□ 4: CID: 16758820

**FIGURE 2** Partial view of an Entrez PubChem Compound display showing the 19 results having the same connectivity as CID 13342.



BioActivity Analysis: 341 Bioassays and 46 Compounds (7 Tested)

Total Pages: 18 Display: 20

| # | □ | AID | Active | Inactive | Discrepant | Tested | BioAssay | Protein Target |
|---|---|---|---|---|---|---|---|---|
| 1 | □ | 1 | 1 |  |  | 1 | NCI human tumor cell line growth inhibition assay. Data for the NCI-H23 Non-Small Cell Lung cell line[Confirmatory] |  |
| 2 | □ | 3 | 1 |  |  | 1 | NCI human tumor cell line growth inhibition assay. Data for the NCI-H226 Non-Small Cell Lung cell line[Confirmatory] |  |
| 3 | □ | 5 | 1 |  |  | 1 | NCI human tumor cell line growth inhibition assay. Data for the NCI-H322M Non-Small Cell Lung cell line[Confirmatory] |  |
| 4 | □ | 7 | 1 |  |  | 1 | NCI human tumor cell line growth inhibition assay. Data for the NCI-H460 Non-Small Cell Lung cell line[Confirmatory] |  |
| 5 | □ | 9 | 1 |  |  | 1 | NCI human tumor cell line growth inhibition assay. Data for the HOP-62 Non-Small Cell Lung cell line[Confirmatory] |  |
| 6 | □ | 11 | 1 |  |  | 1 | NCI human tumor cell line growth inhibition assay. Data for the HOP-18 Non-Small Cell Lung cell line[Confirmatory] |  |
| 7 | □ | 13 | 1 |  |  | 1 | NCI human tumor cell line growth inhibition assay. Data for the HOP-92 Non-Small Cell Lung cell line[Confirmatory] |  |
| 8 | □ | 15 | 1 |  |  | 1 | NCI human tumor cell line growth inhibition assay. Data for the NCI-H522 Non-Small Cell Lung cell line[Confirmatory] |  |

**FIGURE 3** Partial view of a PubChem bioactivity analysis display for the 46 compounds that have the same parent compound connectivity to CID 13342. (*See insert for color representation of the figure.*)

of the Entrez result page, we get the analysis summary "341 Bioassays and 46 Compounds (7 Tested)" (see Figure 3), which is consistent with our previous Entrez search. In the default "Summary" tab of this tool, select the active compounds in the box labeled "Revise Compound Selection" and select the active assays in the box labeled "Revise BioAssay Selection." This will narrow the results to four compounds which were tested and found to be active in a total of 107 assays.

The entire list of all 107 assays may be shown by selecting "All" in the "Display" menu near the top. Note that in two assays, AIDs 589 and 590, there are both active and inactive results. In AID 589, CID 5388983 was found to be active, and this is the sulfate form of the original structure (CID 13342) found by the name search. CID 6604041 was found to be inactive and is different from CID 13342 at several stereocenters. It seems reasonable to hypothesize that the difference in activity is due to the difference in stereochemistry, assuming of course that the stereochemistry of these structures was represented correctly in the data supplied.

Going back to the list of 107 active assays, the "Structure Activity" tab of the analysis tool shows a graphical representation of the compounds clustered by structure. In this case the distribution of assay tests is mainly between two structures: CID 241902, tested in 91 assays supplied by NCI, and CID 5388983, in 13 assays from various Molecular Libraries centers. These two compounds (both sulfate forms) differ by a single bridgehead stereocenter, bringing us back to the question of the correct structure of vinblastine. Is this just a mistake in structure representation?

Resolving this question would probably require referring to original chemical literature on the isolation and structure elucidation of this natural product, which is beyond the scope of this work. But it is clear from this example that searching for information in a large database, especially one composed of data from multiple independent sources, is a nontrivial task. This is nothing new to the experienced chemical informatician. An effective search may require some slight fuzziness in textual or chemical structure search parameters to overcome data errors or differences in convention. PubChem has such a large number of "related structure" links for exactly this reason, so that a search may be tailored according to individual needs to make it as precise as necessary but at the same time flexible enough to enable discovery of information that is inexactly related, yet still relevant.

## REFERENCES AND WEBSITES

1. http://pubchem.ncbi.nlm.nih.gov.
2. http://www.chembiogrid.org/related/resources/about.html.
3. *Chimia* **1998**, 52, 652 ff. Thanks to Dr. Kunz of the current *Chimia* editorial team for making this old issue available to us.
4. http://www.rcsb.org/pdb.
5. http://chembiofinderbeta.cambridgesoft.com; Brecher, J. S. *Chimia* **1998**, 52, 658.
6. http://webbook.nist.gov/chemistry.
7. O'Neil, M. J.; Heckelman, P. E.; Koch, C. B.; Roman, K. J., Eds. *The Merck Index: An Encyclopedia of Chemicals, Drugs, and Biologicals*, 14th ed., Merck & Co., Whitehouse Station, NJ, 2006.
8. Weinstein, J. N.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; et al. *Science* **1997**, 275, 343.
9. http://cactvs.nci.nih.gov/ncidb2; Ihlenfeldt, W. D.; Voigt, J. H.; Bienfait, B.; Nicklaus, M. C. *J. Chem. Inf. Comput. Sci*. **2002**, 42, 46.
10. Ertl, P.; Jacob, O. *Theochem* **1997**, 113.
11. http://www.jmol.org; Willighagen, E.; Howard, M. *CDK News* **2005**, 2, 17.

12. Sayle, R.; Milner-White, E. J. *Trends Biochem. Sci*. **1995**, 20, 374.
13. http://mdlchime.com/downloads.
14. http://www.biocheminfo.org/klotho; Kazic, T. *Biosystems* **1999**, 52, 111.
15. http://www.genome.jp/kegg; Ogata, H.; Goto, S.; Fujibuchim, W.; Kanehisa, M. *Biosystems* **1998**, 47, 119.
16. http://www.molecular-networks.com/biopath; Reitz, M.; Sacher, O.; Tarkhov, A.; Trümbach, D.; Gasteiger, J. *Org. Biomol. Chem*. **2004**, 2, 3226.
17. http://www.hmdb.ca; Wishart, D. S.; et al. *Nucleic Acids Res*. **2007**, 35, 521.
18. http://www.brenda-enzymes.info; Schomburg, I.; Chang, A.; Schomburg, D. *Nucleic Acids Res*. **2002**, 30, 47.
19. http://chem.sis.nlm.nih.gov/chemidplus; Wexler, P. *Toxicology* **2004**, 198, 161.
20. http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?TOXLINE Wexler, P. *Toxicology* **2004**, 186, 161.
21. http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?GENETOX; Wexler, P. *Toxicology* **2004**, 186, 161.
22. http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS; Wexler, P. *Toxicology* **2004**, 186, 161.
23. http://chembank.broad.harvard.edu/welcome.htm; Seiler, K. P.; George, G. A.; Happ, M. P.; Bodycombe, N. E.; Carrinski, H. A.; et al. *Nucleic Acids Res*. **2008**, 36, 351.
24. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug. Del. Rev*. **1997**, 23, 3.
25. http://pubchem.ncbi.nlm.nih.gov/pug/pughelp.html.
26. http://www.ebi.ac.uk/chebi/index.jsp.
27. http://www.drugbank.ca; Wishard, D. S.; Knox, C.; Guo, A. C.; et al. *Nucleic Acids Res*. **2006**, 34.
28. http://www.bindingdb.org/bind/index.jsp; Liu, T.; Lin, Y.; Wen, X.; et al. *Nucleic Acids Res*. **2007**, 35.
29. http://sw16.im.med.umich.edu/databases/pdbbind/index.jsp; Wang, R.; Fang, X.; Lu, Y.; Wang, S. *J. Med. Chem*. **2004**, 47, 12.
30. Zhang, J.; Aizawa, M.; Amari, S.; Iwasawa, Y.; et al. *Comp. Biol. Chem*. **2004**, 28, 401.
31. http://www.bindingmoad.org/; Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. *Proteins* **2005**, 60, 333.
32. http://www.inpharmatica.co.uk/StARLITe.
33. http://nihroadmap.nih.gov.
34. http://www.ncbi.nlm.nih.gov/pubmed.
35. http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml; Wang, Y.; Addess, K. J.; Geer, L.; Madej, T.; Marchler-Bauer, A.; Zimmerman, D.; Bryant, S. H. *Nucleic Acids Res*. **2000**, 28, 243.

36. http://www.ncbi.nlm.nih.gov/Genbank; Benson, D. A.; Boguski, M. S.; Lipman, D. J.; Ostell, J.; Oulette, B. F.; Rapp, B. A. M; Wheeler, D. L. *Nucleic Acids Res*. **1999**, 27, 12.
37. http://www.ncbi.nlm.nih.gov/sites/entrez?db=mesh; Humphrey, S. M. *J. Am. Soc. Inf. Sci*. **1984**, 35, 34.
38. http://dailymed.nlm.nih.gov/dailymed/about.cfm.
39. http://pubchem.ncbi.nlm.nih.gov/deposit.
40. http://www.chemspider.com.
41. http://www.mdli.com/downloads/public/ctfile/ctfile.jsp.
42. Weininger, D. *J. Chem. Inf. Comput. Sci*. **1988**, 28, 31–36; http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html.
43. http://www.iupac.org/inchi.
44. http://pubchem.ncbi.nlm.nih.gov/sources.
45. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol*. **1990**, 215, 403.
46. http://blast.ncbi.nlm.nih.gov/Blast.cgi.
47. http://biocyc.org.
48. http://www.epa.gov/ncct/dsstox/index.html.
49. http://dtp.nci.nih.gov.
50. http://chemdb.niaid.nih.gov/struct_search/default.asp.
51. http://www.ncbi.nlm.nih.gov/books/bookres.fcgi/micad/home.html.
52. http://www.nature.com/nchembio.
53. http://nmrshiftdb.ice.mpg.de.
54. http://pubchem.ncbi.nlm.nih.gov/standardize.
55. Ihlenfeldt, W. D.; Gasteiger, J. *J. Comput. Chem*. **1994**, 15, 793.
56. http://pubchem.ncbi.nlm.nih.gov/search.
57. http://www.cas.org/expertise/cascontent/registry/regsys.html.
58. http://www.eyesopen.com/products/toolkits/lexichem-tk_ogham-tk.html.
59. http://www.eyesopen.com/products/toolkits/oechem.html.
60. http://cactus.nci.nih.gov/osra.
61. *J. Chem. Inf. Comput. Sci*., **1992**, 32(4), 373–378.