BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–9

ChemDB: A Public Database of Small Molecules and Related Chemoinformatics Resources

Jonathan Chen, S.Joshua Swamidass*, Yimeng Dou,Jocelyne Bruand, and Pierre Baldi*†

Institute for Genomics and Bioinformatics
School of Information and Computer Sciences
University of California, Irvine

ABSTRACT

Motivation: The development of chemoinformatics has been hampered by the lack of large, publicly available, comprehensive repositories of molecules, in particular of small molecules. Small molecules play a fundamental role in organic chemistry and biology. They can be used as combinatorial building blocks for chemical synthesis, as molecular probes in chemical genomics and systems biology, and for the screening and discovery of new drugs and other useful compounds.

Results: We describe ChemDB, a public database of small molecules available over the Web. ChemDB is built using the digital catalogs of over a hundred vendors and other public sources and is annotated with information derived from these sources as well as from computational methods, such as predicted solubility and 3D structure. It supports multiple molecular formats and is periodically updated, automatically whenever possible. The current version of the database contains approximately 4.1 M commercially available compounds, 8.8 M counting isomers. The database includes a user-friendly graphical interface, chemical reactions capabilities, as well as unique search capabilities.

Availability: Database, datasets, and supplementary materials available through: `http://cdb.ics.uci.edu`.

Contact: pfbaldi@ics.uci.edu

1 INTRODUCTION

The development of chemoinformatics has been greatly hampered by the lack of publicly available, comprehensive, datasets of molecules (Marris, 2005), of large-scale collaborative projects to annotate these molecules, and of efficient tools to rapidly sift through large chemical repositories. Suffice it to say that no repository of all known organic molecules and their properties is publicly available and downloadable over the Internet. To draw a simple analogy with bioinformatics, the chemoinformatics equivalent of GenBank and Blast are still to be created. To begin addressing these problems, at least for organic chemistry, we describe ChemDB, a public database available over the Web and containing over 4.1 million small molecules.

Small molecules with at most a few dozen atoms play a fundamental role in organic chemistry and biology. They can be used as combinatorial building blocks for chemical synthesis (Schreiber, 2000; Agrafiotis *et al.*, 2002), as molecular probes for perturbing and analyzing biological systems in chemical genomics and systems biology (Schreiber, 2003; Stockwell, 2004; Dobson, 2004), and for the screening, design, and discovery of useful compounds. These include of course new drugs (Lipinski and Hopkins, 2004; Jonsdottir *et al.*, 2005), the majority of which are small molecules. Furthermore, huge arrays of new small molecules can be produced in a relatively short period of time (Houghten, 2000; Schreiber, 2000).

As datasets of small molecules become available, it is crucial to organize these datasets in rapidly searchable databases and to develop computational methods to rapidly extract or predict useful information for each molecule, including its physical, chemical, and biological properties. Conversely, large and well-annotated datasets are essential for developing statistical machine learning methods in chemoinformatics, whether supervised or unsupervised, including predictive classification, regression, and clustering of small molecules and their properties (e.g. Micheli *et al.* (2003); Ralaivola *et al.* (2005)). Aggregation and organization of datasets of chemical information allows for massive in silico processing that would be impractical or even impossible in a traditional experimental setting.

Consider, for instance, a classical drug discovery problem where the starting point is a protein of known structure and perhaps a corresponding ligand (Figure 1). With a good database of small molecules, the discovery process can proceed from both ends. Starting from the protein, one can dock millions of small molecules to the protein in silico. In fact, with sufficient computing power, one ought to be able to dock all known small molecules to all proteins with known structure

*These authors contributed equally.

†To whom correspondence should be addressed.

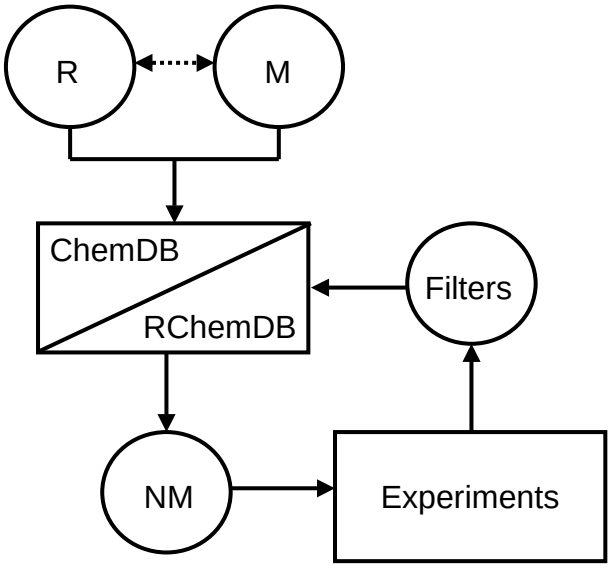

Fig. 1. High-level view of a basic drug screening/design pipeline. R = receptor protein; M = molecular ligand(s); NM = new molecular ligands; RChemDB = set of compounds derived from ChemDB using a library of reactions. NM is obtained by molecular docking applied to R, or by constrained similarity searches applied to M. Computational filters can be used to predict and constrain molecular properties (e.g. flexibility, solubility, toxicity).

contained in the PDB (Berman *et al.*, 2000). Starting from the ligand, one can search the database of small molecules for compounds that are "similar" to the known ligand(s), where similarity can be defined in different ways. In both approaches, additional filters can be used to eliminate molecules that are, for instance, poorly soluble, too flexible, or toxic (Swamidass *et al.*, 2005). Furthermore, in silico chemical reactions applied to the molecules in the database can further expand the space of interesting molecules being screened or designed.

Most large databases of small molecules, such as MDL's Available Chemicals Directory (ACD) or American Chemical Society's CAS registry, are privately owned, expensive, and often available only through restricted interfaces that are not suitable for the development of statistical methods. A few datasets of small molecules, such as the NCI (National Cancer Institute) open database, are available publicly. However, in general these are limited in size, with on the order of 10^3 to 10^5 compounds (Voigt *et al.*, 2001). Furthermore, efforts towards public databases must face fierce opposition from the ACS (Marris, 2005; Kaiser, 2005*a*,*b*). Given the importance of small molecules, ChemDB aims to address the data bottleneck in the current environment by integrating existing public datasets with datasets originating from dozens of chemical vendors. These datasets are integrated into a database containing on the order of 10^7 compounds, available over the Web with a unique combination of chemoinformatics resources.

2 METHODS

2.1 Data Sources, Formats, and Size

ChemDB is a chemical information database system grown out of an aggregation of multiple information sources, primarily commercial vendor catalogs, but also publicly available repositories (e.g. NCI). For vendors and other sources that periodically update their data and make them available over the Internet, we automatically download the data and resynchronize the latest updates into ChemDB. For other sources that currently distribute their data only through CDs, we contact them periodically for updates. Complete information about all the vendors is available from the supplementary materials. In total, the current database contains about 4.1 M unique compounds, 8.8 M counting isomers, aggregated from over 115 sources.

Molecules come with multiple representations and formats (Figure 2) including 1D SMILES strings (Weininger *et al.*, 1989; James *et al.*, 2004), 2D graphs of atoms and bonds, 3D atom coordinates (SDF or MOL2 files), and fingerprints (Fligner *et al.*, 2002; Flower, 1998; Ralaivola *et al.*, 2005), all of which are stored in ChemDB. We have developed scripts to automatically parse input data, run different tests, and populate the database. To populate the database, all datasets are first converted to the SDF format because of its standardized annotation mechanism. However, conversion between several popular molecular file formats, including SMILES, MDL Mol, PDB, Tripos Sybyl mol2, and SDF is easily accomplished using OpenEye Software's OEChem toolkit (`http://www.eyesopen.com`), or the open-source alternative, OpenBabel. Additional curation and normalization steps are applied to the data as it is inserted. For instance, 3D structures are generally not available and are therefore predicted using the program CORINA (Sadowski *et al.*, 1994; Gasteiger *et al.*, 1996).

One difficult issue for chemoinformatics systems is how to handle stereochemistry. This issue is complicated by the absence of stereochemical and geometric information from most sources, which generally provide only the atom-bond connection table. ChemDB currently enumerates up to $n = 16$ stereoisomers for each molecule. This is a reasonable number since it allows listing *all* possible isomers for over 97.4% of compounds in ChemDB, i.e. those with at most 4 stereocenters and therefore at most 16 isomers (see Results). In addition, for each isomer, ChemDB generates and stores not only the stereochemistry specific connection table as an isomeric SMILES string, but also the corresponding predicted 3D coordinates as an SDF file. This solution allows us to specify which isomer is relevant when stereochemistry is known and provides a more complete picture for the user, in virtual docking and other applications. If the stereoisomer is not specified, we assume that the chemical is available as a racemic mixture. Thus, this solution provides a reasonable and

effective compromise in light of limited information about molecule handedness.

In the future it may be possible to heuristically guess more relevant isomers by cross-referencing structures with other database such as the PDB PubChem and the PDB and perhaps more intelligent decoding of chemical names: our database schema immediately accommodates these extensions.

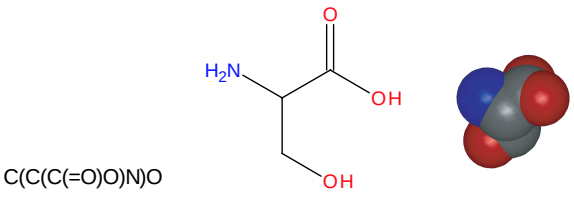

Fig. 2. Three representations for the amino acid Serine: (a) 1D SMILES string; (b) 2D graph of atoms and bonds; and (c) 3D space-fi lling model. A fourth representation based on fi ngerprint vectors is briefly described in the text.

2.2 Database Schema

The basic database schema is relationally organized and relies upon canonical SMILES string representations for rapid indexing and enforcement of uniqueness. The relational structure allows for maintenance and querying of complex arrangements, such as the many-to-many relationship between sources and the chemicals for which they provide records. The database schema contains primary tables for sources, chemicals, and molecular descriptors and annotations. It is described in detail in the supplementary materials.

2.3 Implementation

The database is implemented using the leading open-source relational database PostgreSQL (`http://www.postgresql.org`). We have also built filters for conversion to Oracle and maintain an Oracle version internally for comparison purposes. Web interfaces and tools are delivered using the open-source Apache Web server. Many of the basic application tools, scripts and Web interfaces are written in Python, while computationally intensive modules are written in C or Java. Python has convenient interfaces to important packages, like the OEChem toolkit which implements several basic algorithms needed for chemical data processing, including SMARTS pattern matching and SMIRKS reaction processing. We use OEDepict and the JMol Java applet (`http://jmol.sourceforge.net/`) for chemical image rendering.

2.4 Molecular Descriptors and Example of Filters

In addition to 3D structures obtained using CORINA, we compute and store several other molecular descriptors including: molecular weight, number of hydrogen-bond donors, number of hydrogen-bond acceptors, octanol/water partition coefficient $\log P$, solvation energy, number of rigid fragments, number of rotatable bonds, number of chiral centers, and number of chiral double bonds. For each molecular descriptor, we include hyperlinks to the program that was used to compute it. For instance, we compute $\log P$ values for all compounds, using the XLogP program and the calculation module available from ChemAxon (`http://www.chemaxon.com`). Similarly, compound solvation energy is always calculated and recorded using OpenEye's ZAP module. We also store in a similar way any additional molecular descriptors found in the vendor 's electronic catalogs.

The database interface allow the user to implement flexible search filters by specifying thresholds or ranges for any combination of these molecular descriptors. For example, Lipinski's rules of five (Lipinski *et al.*, 1997) are often used as criteria for drug oral bioavailability. These rules correspond to molecular mass less than 500 daltons; number of hydrogen-bond donors less than 5; number of hydrogen-bond acceptors less than 10; and octanol/water partition coefficient $\log P$ (an indication of the ability of a molecule to cross biological membranes) less than 5. If two or more of those criteria are out of range, the compound is likely to have poor absorption or permeability. The cutoffs in the rules can be tightened or relaxed in the interface to allow for flexible searches as well as computational and experimental errors, especially in the computational determination of the partition coefficient. As an alternative, one can easily use the set of rules proposed in Veber *et al.* (2002). By examining oral bioavailability in rats for over 1100 drug candidates, Veber et al. concluded that only two structural variables control this crucial property: molecular flexibility, measured by the number of rotatable bonds, and polar surface area, expressed as the sum of hydrogen bond donors and acceptors. These studies indicate that drug candidates with 10 or fewer rotatable bonds and a polar surface area equal to or less $140 Å^2$ (equivalent to 12 or fewer H-bond donors and acceptors) will exhibit favorable oral bioavailability.

It is important to recognize, however, that the Lipinski and Veber rules are not absolute (Frimurer *et al.*, 2000) and that oral bioavailability is only one of many potentially important criteria. New modes of drug delivery have entered clinical practice recently and will likely continue to do so in the future. Thus, even within the limited chemoinformatics goals of drug screening, the ChemDB interface provides the user with a wide array of filters and threshold values to be tailored to different problems and searches.

2.5 Vendor/Source Descriptors and Experimental Annotations

We store the name, contact information, and date of the latest update for each vendor's dataset in the database. In addition, we incorporate any annotation provided in the vendor's digital catalogs. These annotations range in utility and their presence

can vary greatly from vendor to vendor. Annotations provided typically include: purchase price, English name, CAS registry number, experimentally determined octanol-water partition coefficient, amount available, purity, melting point, heteroatoms, and net charge. For a smaller fraction of compounds, additional miscellaneous information is available from the vendors, ranging from literature references, to boiling points, to possible activity (e.g. Interleukin agonist). We flag any experimentally derived annotation provided by the vendors or present in some of the public datasets (e.g. NCI). Finally, we also flag FDA-approved drugs for reference.

2.6 Similarity Search Methods and Kernels

As in the case of bioinformatics, once large repositories of small molecules are assembled, the next fundamental step for chemoinformatics is the definition and fast implementation of similarity measures. This is fundamental for two reasons: (1) to enable rapid and meaningful searches through million of records; and (2) to enable supervised and unsupervised predictive methods that are based on similarity measures, from clustering to kernel methods (Schölkopf and Smola, 2002).

Similarity measures between small molecules can be defined in several different ways and by leveraging different representations. In Swamidass *et al.* (2005), several similarity measures are described and assessed using spectral representations and spectral kernels, i.e. similarity measures derived by comparing the occurrence of substructures, such as substrings in SMILES strings, paths in 2D atom-bond graphs, and histograms of atomic distances in 3D. While these and other similarity measures are under investigation, at this point the 2D similarity measures yield the best results and are used extensively in ChemDB. These measures are based on fixed-size fingerprint vectors counting the presence (or number of occurrences) of labeled paths in a molecule (see Swamidass *et al.* (2005) and references therein for details). Binary fingerprint representations of typical length 512 or 1024, combined with efficient bit-wise algorithms, yield fast search algorithms.

2.7 Web Interface

The speed with which the bit-wise algorithm can sequentially search millions of chemical fingerprints makes the ChemDB available for queries through a Web interface, where users can enter query molecule(s) in any standard molecular file format with several additional options. These novel options include the ability to search by a selection of ranges for the primary annotations (e.g. number of rotatable bonds), by substructures and superstructures with or without constraints on the presence or absence of particular groups or other features (masks), and by "profile", using a group of related molecules rather than a single molecule as the query. The current version of profile search maximizes the sum of the similarities to each of the query structures. Alternatives under investigation include building a profile fingerprint vector and using alternative measures, such as the maximum of the pairwise similarities between a molecule and the structures in the query set.

2.8 In Silico Chemical Reactions: RChemDB

The repository's size can be expanded further by considering virtual compounds that can be synthesized from building-blocks in the ChemDB, which are readily available through the vendors. This can be achieved by annotating functional groups and applying in silico reactions to the current dataset. Implicit or explicit functional group annotation is derived using the SMARTS pattern method (James *et al.*, 2004) with the OEChem implementation. The SMARTS pattern method is essentially a subgraph isomorphism algorithm applied to a molecule represented as a graph of labeled atoms and bonds. This method provides precise search results, but is computationally intensive and therefore not suitable for interactive use. In comparison, the fingerprint bit vector approach may have a slightly higher rate of false positives but is much faster and therefore more suitable for interactive use. Furthermore, unlike a simple SMARTS-based approach which only provides a binary result–depending on whether a given substructure is present or not in a given structure–a fingerprint-based approach provides a real-valued similarity score between any two structures.

Once functional groups have been identified, combinatorial reactions that specify which groups can react are defined by the Daylight SMIRKS specification (James *et al.*, 2004). Examples of reactions currently implemented include: Amide Formation, Buchwald-Hartwig, Cyanation of Aromatic Halides, DielsAlder, Ester Formation, Grignard, GrubsReaction, Heck, Hiyama, Negishi, Phosphodiester Formation, Sonogashira, Suzuki, and SwernOxidation. RChemDB denotes the set of virtual molecules that can be generated from ChemDB by iterative applications of a library of in silico reactions. It is essential to note that as the reactions are iterated, the number of compounds grows exponentially. Thus RChemDB itself is virtual in the sense that we can generate and conduct directed searches of its compounds, but these are not stored directly into ChemDB. An example of application of RChemDB is small chemical refinement of a basic lead or scaffold, by letting the scaffold structure react with all of the very small molecules in ChemDB, e.g. with less than 10 atoms. These correspond to slightly less than 1% of ChemDB.

2.9 Additional Datasets

In addition to particular subsets that can be extracted from ChemDB, we also maintain on a Web page associated with the ChemDB a list of downloadable datasets that can be used as training/validation sets in unsupervised or supervised machine learning and other computational experiments. These are also hyperlinked with the UCI Machine Learning Repository.

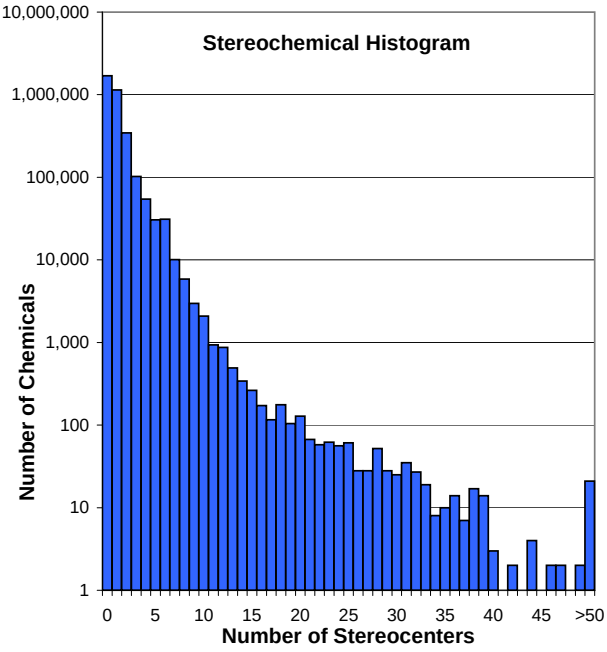

Fig. 3. Histogram of number of stereocenters per chemical in ChemDB on a log scale. The fraction of compounds with at most 16 isomers (4 stereocenters) is 97.5%. All entries with 50 or more stereocenters are binned at 50.

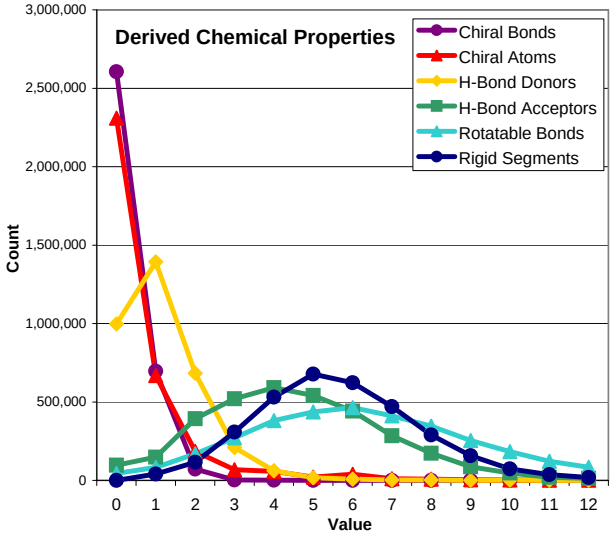

Fig. 4. Histogram of several molecular descriptors calculated for all chemicals in ChemDB (unique records). Properties covered here include chiral atoms and chiral bonds to identify structures that can exhibit stereochemistry, rotatable bonds and rigid segments as a measure of molecular flexibility, and H-bond donors and H-bond acceptors based on Lipinski's definitions of hydrogens attached to a nitrogen or oxygen (donor) and any nitrogen or oxygen (acceptor).

3 RESULTS

3.1 Statistics

ChemDB allows us to compute several useful statistics on small molecules. For example, the histogram in Figure 3 counts the number of molecules in ChemDB with a given number of stereocenters. A molecule with k stereocenters generally yields 2^k isomers or less. This is because isomers are based on stereocenters that generally have two configurations. The number of isomers can be less than 2^k due to geometric clashes or redundant combinations of stereocenters. The majority of chemicals in the system (2.5M) have no stereocenters, which explains why although CORINA is set to determine up to 16 isomers per chemical, the number of isomers (7.4 M) is only about twice the number of unique records. In fact, 97.4% of the chemicals in ChemDB have at most 4 stereocenters, resulting in at most 16 configurations, and thus all their isomers are stored in ChemDB. Using values $k = 5$ or $k = 6$ would increase the coverage only marginally to 98.3% and 99.2% respectively. For the small minority of compounds that have more than 4 stereocenters, rather than pre-computing and storing all configurations, a random sample of 16 isomers is precomputed and stored in ChemDB. Additional isomers can be generated on a per-request-basis. Finally, at the extreme end of the distribution, the last bin of the histogram in Figure 3 represents all chemicals with 50 or more stereocenters, of which there are 27 in the database. The top 4 (Cyanovirin, Heptakis-(6-O-maltosyl)-β-cyclodextrin, D-Alanyl-lipoteichoic acid, and Scytovirin) have 111, 105, 86, and 86 stereocenters, respectively. Enumerating all of the isomers for the first chemical alone would yield potentially $2^{111} \approx 10^{33}$ possible configurations. It is worth noting, however, that the majority of the chemicals in this tail are natural products, or natural product derivatives. Thus only one or two isomers of each chemical are likely to exist in nature and be available from vendors.

ChemDB histograms for other molecular descriptors including the number of chiral bonds, chiral atoms, H-bond donors, H-bond acceptors, rotatable bonds, and rigid segments per molecule are shown in Figure 4. Additional pairwise statistics, displaying for instance the weak correlation betwen molecular weight and (predicted) solubility, are given in the supplementary materials.

3.2 In Silico Reactions

Examples of simple reaction processing capabilities implemented in ChemDB are given in Figures 5 and 6. Figure 5, derived from the ChemDB interface, shows how amino groups and carboxylic acids react to form an amide bond. Besides expanding the data set by predicting reaction products, the reaction processing capabilities can be applied to other novel purposes. For example, they can be used as part of a screen for potential polymer components. A simple polymer screen–identifying candidates that can at least self-polymerize–is accomplished by identifying each molecule which, given a library of reactions, can iteratively react with itself and with

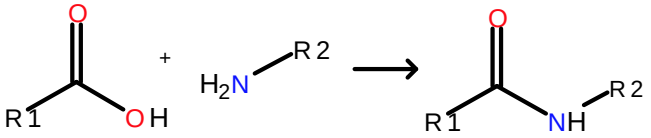

Fig. 5. SMIRKS reaction specification (Carboxylic acid + amine >> amide): Depiction of SMIRKS reaction for amide formation: [O:1]=[C:2][O:3][H:7].[H:8][N:4][H:5]>>[O:1]=[C:2][N:4][H:8].

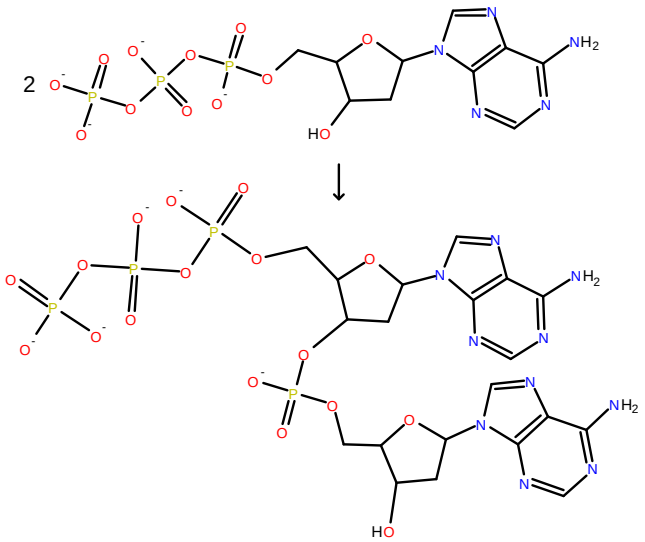

Fig. 6. dATP and di-dATP as an example of using reaction processing tools to identify polymer candidates. dATP passes the simple polymer screen because it can react with itself to yield a product with the same properties, forming the initial components of a DNA polymer.

the products of these reactions. Figure 6 shows how DNA can be "rediscovered" in ChemDB using a simple polymer screen.

3.3 Web Interface and Searches

Figure 7 depicts a composite screenshot from the ChemDB interface upon performing an integrated chemical similarity search. Shown inset on the left is the structure for a chemical known to be an inhibitor of monoacylglycerol lipase (MGL), an intracellular serine hydrolase that catalyzes the hydrolysis of 2-arachidonoylglycerol (2-AG), a primary endogenous cannabinoid in the mammalian brain. Recent studies suggest an MGL inhibitor can mediate opioid-independent stress-induced analgesia, identifying MGL as an important drug target (Hohmann *et al.*, 2005). In an effort to find additional inhibitors in collaboration with chemists and pharmacologists (Drs. Chamberlin and Piomelli), ChemDB was searched for chemicals with structural similarity to a known inhibitor. In this case similarity is computed using the Tversky measure (Tversky, 1977; Rouvray, 1992) applied to the binary fingerprints. Based upon a mechanistic understanding of the inhibitor, our collaborators provided feedback suggesting that chemicals of interest may not require complete structural similarity to the original chemical. Instead, the two structures shown in the sketcher window in the top left, should be contained as sub-structures/functional groups of the desired structure. As shown in the figure, beyond substructures, the search can be further refined by restricting the ranges of several molecular descriptors, in this case the number of rotatable bonds, predicted XLogP, and molecular weight. The particular set of values selected in this example reflect a customized combination of Lipinski's and Veber's rules. All compounds are ranked in decreasing order of similarity, but only the top 3 results are shown here together with their similarity score and basic information, including corresponding vendor. The interface also displays a dynamically generated histogram representing the similarity score distribution for every chemical in the database relative to the original structure on a logarithmic scale. After human-expert examination, several top hits obtained from these searches have been ordered from the corresponding vendors and are being tested in the laboratory.

4 DISCUSSION

While most commercial databases of small molecules have sizes smaller or comparable to ChemDB, there exist a few that are larger, notably the CAS registry of the American Chemical Society (ACS) with the related SciFinder tool. While these commercial databases may contain useful information, they do not always provide flexible chemoinformatics tools or interfaces. For instance, even in the ACS database, queries are allowed only one compound at a time and the full database is not downloadable. As in bioinformatics with Genbank or PDB, queries performed one item at a time may be satisfactory for many users, but researchers involved in the development and application of large-scale datamining methods need full access to the entire corpus of data. Furthermore, the cost of these commercial databases is often very significant, at least from an academic standpoint.

To address the data bottleneck created in part by the ACS, public, downloadable, chemical repositories have begun to emerge. In addition to NIH's PubChem, (`http://pubchem.ncbi.nlm.nih.gov`), examples of other public database efforts related to ChemDB include Harvard's ChemBank (Strauseberg and Schreiber, 2003), UCSF's ZINC (Irwin and Shoichet, 2005), and the European Bioinformatics Institute's ChEBi (`http://www.ebi.ac.uk/chebi`). While in the long run some degree of consolidation among these efforts can be expected–we are currently depositing ChemDB compounds into PubChem–in the short run a diversity of efforts with different aims and approaches allows the

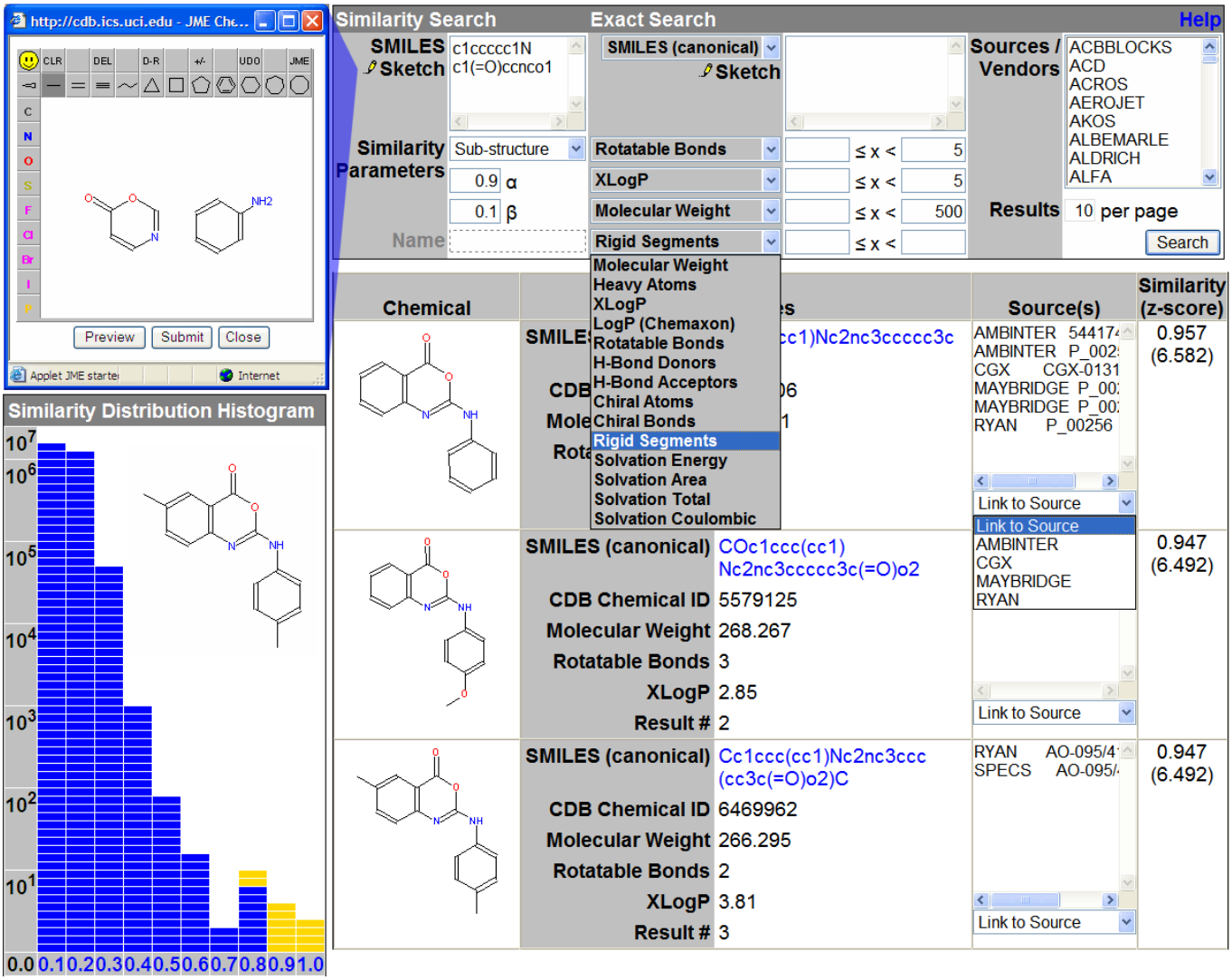

Fig. 7. Composite screenshot example of an integrated search. The two structures shown in the sketcher window in the top left are simultaneously considered in a similarity search, but only as sub-structures/functional groups. The substructure bias is accomplished by setting the alpha and beta parameters of the Tversky similarity measure to 0.9 and 0.1, respectively. Results are ranked by similarity score, ranging from 0.0 to 1.0. Below each raw similarity score is the corresponding z-score (i.e. the number of standard deviations away from the mean of similarity scores for all known chemicals in the database). The similarity search is integrated with standard fi lters like those shown to restrict the results by number of rotatable bonds, predicted XLogP, and molecular weight. More generally, using the drop-down menu, the user can specify ranges for combinations of 14 molecular descriptors (H-Bond Acceptors, H-Bond Donors, Molecular Weight, LogP, XLogP, Heavy Atoms, Rotatable Bonds, Chiral Atoms, Chiral Bond, Rigid Segments, Solvation Energy, Solvation Area, Solvation Total, Solvation Coulombic). A few top results for this search are shown. Shown on the left is a dynamically generated histogram representing the similarity score distribution for every chemical in the database, with respect to the original search structure shown inset with the histogram. Note that the histogram is displayed on a logarithmic scale and that the blocks corresponding to the fi rst"page" worth of results (top 10) are automatically highlighted.

exploration of different solutions and tradeoffs. Indeed, the existing databases have slightly different goals and properties, in terms of size, focus, availability, and informatics algorithms for searching and other operations. At the time of this writing, for instance, Pub Chem and ChemBank are smaller in size (approximately 1 M compounds) with a greater emphasis on literature references (PubChem) and experimental bioactivity annotation (Chembank). Unlike the other public databases, PubChem and ChemBank allow also searching compounds

by English names. ChemBank, however, is not fully downloadable. ZINC is fully downloadable and perhaps closest in size and spirit to ChemDB, with a primary focus on structure download to facilitate docking. Unlike ZINC and other public repositories, ChemDB's focus goes beyond drug discovery and includes the development of new computational tools for annotating, searching, and mining large repositories of chemical data. In particular, the current flexible search capabilities found in ChemDB are unique and so are its chemical reaction capabilities, among publicly available databases. A table in the supplementary materials summarizes some of the tradeoffs between these synergistic public efforts. Such a table, however, quickly becomes outdated as all of these repositories are undergoing rapid evolution.

Chemical descriptors and annotations are clearly essential for exploring chemical space directly, as well indirectly for the development of efficient computational annotation methods. Many molecular descriptors described here, such as molecular weight and number of rotatable bonds, are precisely defined and can be computed exactly. Other computational annotations, such as the degree of solubility ($\log P$) or 3D structures, are noisy and subject to closer scrutiny. In particular, the predicted 3D structures are important but of some concern since we, and other authors, have noted that predictive methods based on 3D structure can be outperformed by methods based on 2D structure alone (Swamidass *et al.*, 2005). While predicting the structure of small molecules is easier than predicting the structure of proteins, it is still essential to run large-scale tests to assess the quality of those predictions and whether they can be used reliably. For this purpose, we have recently acquired a license to the Cambridge Structural Database system, another commercial repository, containing the experimentally determined 3D structures of about 300,000 molecules to validate the quality of predicted structures.

A related important problem arises with stereochemistry. In ChemDB we have adopted the solution of storing up to 16 isomers for each compound, but we currently do not test for the relevance or synthetic feasibility of these compounds. In the future it may be possible to heuristically guess more relevant isomers by cross-referencing structures with other database such as PubChem and the PDB and perhaps more intelligent decoding of chemical names: our database schema immediately accommodates these extensions. As far as synthetic feasibility, it is an important criterion that should also be implemented in the future. Currently, we believe the enumeration of theoretical compounds has value in and of itself, as we are not just interested in cataloging known compounds but also pushing the boundaries of knowledge towards potential compounds for a better understanding of chemical space. Even from a practical standpoint, a theoretical compound found to be of particular value in a docking study, for instance, may spur the interest of chemists towards its synthesis. Without enumeration of theoretical compounds, this particular compound would have not even been considered. Furthermore, these theoretical compounds are not random but logically derived by computational methods that stack the odds in favor of finding a reasonable synthetic pathway.

Finally, the scarcity of publicly available chemical annotation points to the need for new approaches to chemical annotation that could include: (1) development of automated information retrieval systems to derive annotations from chemical literature; (2) sharing of private or commercial annotation by, for instance, the ACS or large pharmaceutical companies; and (3) development of collaborative, coordinated, and large-scale annotation efforts across academic centers, similar to those used in the other life sciences. Coupling public databases with public annotation efforts will lead in time to repositories that may allow predictive chemical informatics to blossom and develop tools to fully explore chemical space, from drug discovery, to new materials, to the origin of life.

ACKNOWLEDGMENT

Work supported by an NIH Biomedical Informatics Training grant (LM-07443-01) and an NSF MRI grant (EIA-0321390) to PB, by the UCI Medical Scientist Training Program, and by a Harvey Fellowship to S.J.S. We would like also to acknowledge the OpenBabel project and OpenEye Scientific Software for their free software academic license, and Drs. Chamberlin, Nowick, Piomelli, and Weiss for their useful feedback.

REFERENCES

Agrafi otis,D.K., Lobanov,V.S. and Salemme,F.R. (2002) Combinatorial informatics in the post-genomics era. *Nature Reviews Drug Discovery,* **1**, 337–346.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucl. Acids Res.,* **28**, 235–242.

Dobson,C.M. (2004) Chemical space and biology. *Nature,* **432**, 824–828.

Fligner,M.A., Verducci,J.S. and Blower,P.E. (2002) A modifi cation of the Jaccard/Tanimoto similarity index for diverse selection of chemical compounds using binary strings. *Technometrics,* **44** (2), 1–10.

Flower,D.R. (1998) On the properties of bit string-based measures of chemical similarity. *J. of Chemical Information and Computer Science,* **38**, 378–386.

Frimurer,T.M., Bywater,R., Naerum,L., Lauritsen,L.N. and Brunak,S. (2000) Improving the odds in discriminating "drug-like" from "non drug-like" compounds. *Journal of Chemical Information and Computer Sciences,* **40**, 1315–1324.

Gasteiger,J., Sadowski,J., Schuur,J., Selzer,P., Steinhauer,L. and Steinhauer,V. (1996) Chemical information in 3D-space. *Journal of Chemical Information and Computer Sciences,* **36**, 1030–1037.

Hohmann,A.G., Suplita,R.L., Bolton,N.M., Neely,M.H., Fegley,D., Mangieri,R., Frey,J.K., Walker,J.M., Holmes,P.V., Crystal,J.D., Duranti,A., Tontini,A., Mor,M., Tarzia,G. and Piomelli,D. (2005) An endocannabinoid mechanism for stress-induced analgesia. *Nature,* **435**, 1108–1112.

Houghten,R.A. (2000) Parallel array and mixture-based synthetic combinatorial chemistry: tools for the next millenium. *Annual Review of Pharmacology and Toxicology,* **40**, 273–282.

Irwin,J.J. and Shoichet,B.K. (2005) ZINC–a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Computer Sciences,* **45**, 177–182.

James,C.A., Weininger,D. and Delany,J. (2004) *Daylight Theory Manual*. Available at http://www.daylight.com/dayhtml/doc/theory/theory.toc.html.

Jonsdottir,S.O., Jorgensen,F.S. and Brunak,S. (2005) Prediction methods and databases within chemoinformatics: Emphasis on drugs and drug candidates. *Bioinformatics,* **21**, 2145–2160.

Kaiser,J. (2005*a*) Chemists want NIH to curtail database. *Science,* **308**, 774.

Kaiser,J. (2005*b*) House approves 0.5% raise for NIH, comments on database. *Science,* **308**, 1729.

Lipinski,C. and Hopkins,A. (2004) Navigating chemical space for biology and medicine. *Nature,* **432**, 855–861.

Lipinski,C.A., Lombardo,E., Dominy,B.W. and Feeney,P.J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews,* **23** (3), 3–25.

Marris,E. (2005) Chemistry society goes head to head with NIH in fight over public database. *Nature,* **435** (7043), 718–719.

Micheli,A., Sperduti,A., Starita,A. and Biancucci,A.M. (2003) A novel approach to QSPR/QSAR based on neural networks for structures. In *Soft Computing Approaches in Chemistry*, (Cartwright,H. and Sztandera,L.M., eds),. Springer Verlag Heidelberg, Germany pp. 265–296.

Ralaivola,L., Swamidass,S.J., Saigo,H. and Baldi,P. (2005) Graph kernels for chemical informatics. *Neural Networks,* . Special issue on Neural Networks and Kernel Methods for Structured Domains. In press.

Rouvray,D. (1992) Definition and role of similarity concepts in the chemical and physical sciences. *Journal of Chemical Information and Computer Sciences,* **32** (6), 580–586.

Sadowski,J., Gasteiger,J. and Klebe,G. (1994) Comparison of automatic three-dimensional model builders using 639 X-ray structures. *Journal of Chemical Information and Computer Sciences,* **34**, 1000–1008.

Schölkopf,B. and Smola,A.J. (2002) *Learning with Kernels, Support Vector Machines, Regularization, Optimization and Beyond*. MIT University Press.

Schreiber,S.L. (2000) Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science,* **287**, 1964–1969.

Schreiber,S.L. (2003) The small-molecule approach to biology: chemical genetics and diversity-oriented organic synthesis make possible the systematic exploration of biology. *Chemical and Engineering News,* **81**, 51–61.

Stockwell,B.R. (2004) Exploring biology with small organic molecules. *Nature,* **432**, 846–854.

Strauseberg,R.L. and Schreiber,S.L. (2003) From knowing to controlling: a path from genomics to drugs using small molecule probes. *Science,* **300** (5617), 294–295.

Swamidass,S.J., Chen,J., Bruand,J., Phung,P., Ralaivola,L. and Baldi,P. (2005) Kernels for small molecules and the prediction of mutagenicity, toxicity, and anti-cancer activity. *Bioinformatics,* **21** (Supplement 1), i359–368. Proceedings of the 2005 ISMB Conference.

Tversky,A. (1977) Features of similarity. *Psychological Review,* **84** (4), 327–352.

Veber,D., Johnson,S.R., Cheng,H., Smith,B.R., Ward,K.W. and Kopple,K.D. (2002) Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry,* **45**, 2615–2623.

Voigt,J.H., Bienfait,B., Wang,S. and Nicklaus,M.C. (2001) Comparison of the NCI open database with seven large chemical structural databases. *Journal of Chemical Information and Computer Sciences,* **41** (3), 702–712.

Weininger,D., Weininger,A. and Weininger,J.L. (1989) SMILES. 2. algorithm for generation of uniques SMILES notation. *Journal of Chemical Information and Computer Sciences,* **29**, 97–101.