BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 17 2007, pages 2348–2351
doi:10.1093/bioinformatics/btm341

Databases and ontologies

ChemDB update—full-text search and virtual chemical space

Jonathan H. Chen, Erik Linstead, S. Joshua Swamidass, Dennis Wang and Pierre Baldi*

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

Received on April 4, 2007; revised on June 4, 2007; accepted on June 21, 2007

Advance Access publication June 28, 2007

Associate Editor: Chris Stoeckert

ABSTRACT

ChemDB is a chemical database containing nearly 5M commercially available small molecules, important for use as synthetic building blocks, probes in systems biology and as leads for the discovery of drugs and other useful compounds. The data is publicly available over the web for download and for targeted searches using a variety of powerful methods. The chemical data includes predicted or experimentally determined physicochemical properties, such as 3D structure, melting temperature and solubility. Recent developments include optimization of chemical structure (and substructure) retrieval algorithms, enabling full database searches in less than a second. A text-based search engine allows efficient searching of compounds based on over 65M annotations from over 150 vendors. When searching for chemicals by name, fuzzy text matching capabilities yield productive results even when the correct spelling of a chemical name is unknown, taking advantage of both systematic and common names. Finally, built in reaction models enable searches through virtual chemical space, consisting of hypothetical products readily synthesizable from the building blocks in ChemDB.

Availability: ChemDB and Supplementary Materials are available at http://cdb.ics.uci.edu.

Contact: pfbaldi@ics.uci.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Small organic molecules play an increasingly important role in science, from building blocks in chemical synthesis, to molecular probes in systems biology, to therapeutic drugs in medicine. Until recently, the large-scale study of small molecules has been hampered by the lack of comprehensive, publicly available, datasets and collaborative projects to annotate them. To draw a simple analogy, the chemoinformatics equivalent of GenBank is still to be established. A few systems have begun to address this need, including PubChem (http://pubchem.ncbi.nlm.nih.gov), ZINC (Irwin and Shoichet, 2005), ChEBi (http://www.ebi.ac.uk/chebi), Chembank (Strauseberg and Schreiber, 2003) and ChemDB (Chen *et al*., 2005).

*To whom correspondence should be addressed.

2 DATABASE DESCRIPTION

The underlying data in ChemDB primarily comes from the electronic catalogs of over 150 chemical vendors as well as a limited number of publicly available datasets, such as the NCI small molecule screening library (Voigt *et al*., 2001). The chemical structures are stored in the database together with a number of calculated descriptors, such as molecular weight and H-bond donors, as well as predicted properties, such as 3D structure. More recently, several new physicochemical property predictors have been added internally to ChemDB. The underlying predictors are also available externally in standalone mode through an online form enabling predictions for arbitrary input molecules. Several of these predictors are based on machine-learning techniques including kernel methods: kLogP (octanol/water partition coefficient), kSol (aqueous solubility) and kMelt (melting point) (Azencott *et al*., 2007).

3 DATABASE SEARCH OPTIONS

Chemoinformatics data repositories with millions of records require efficient, 'BLAST-like', search methods to sift through the data and retrieve useful results. ChemDB includes several new search tools in support of systems biology and drug discovery projects. These are illustrated here using examples from an ongoing tuberculosis (TB) drug discovery project (Lin *et al*., 2006).

3.1 Searching by structural similarity

In this project, a fatty acid biosynthesis enzyme specific to *Mycobacterium tuberculosis* was identified as a possible therapeutic target. Solution of the enzyme 3D structure by X-ray diffraction enables the search for inhibitors through repeated cycles of *in silico* docking and bioassays. At each cycle, and more generally whenever a lead becomes available, the search can be expanded by looking for small molecules in ChemDB that are 'similar' to the lead.

The original release of ChemDB included a chemical fingerprint-based method to search for similar chemicals based on atom-bond connectivity. The search algorithm has gone through several upgrades, including a mathematical correction which provides better estimates of uncompressed fingerprint similarity starting from the compressed fingerprints, thereby improving retrieval accuracy (Swamidass and Baldi, 2007b). Speed optimization also now yields sublinear search

times (Swamidass and Baldi, 2007a). The optimization is based on organizing chemical fingerprints by bit density and dynamically pruning the database during a given search by excluding compounds outside the relevant density range. In practice, searches across 5M molecules that used to take up to 30 s, can now be completed in less than a second on a single computer.

Among the more unique features of ChemDB's structural search engine is its ability to search for super- and sub-structures (e.g. functional groups) and to perform profile searches, using multiple molecules in the query. In combination, sub-structure and profile searches enable one to quickly find molecules that contain several key functional groups. Collectively, these search methods have continued to provide fruitful leads (Lin *et al*., 2006).

3.2 Searching by names and annotations

In reference to pre-existing treatments for tuberculosis, a researcher may be interested in retrieving information, such as 3D structures for docking studies, on the drugs isoniazid and rifampicin. However, without prior knowledge of the chemical content or structure of these drugs, the researcher would have been unable to find these records. ChemDB now allows searching for compounds by name and other text annotations.

Chemical vendors often supply text annotations and descriptors in their catalogs, including common and systematic names, bioactivity assay results and CAS numbers. These annotations, numbering over 65M, comprise a substantial corpus of information. To allow users to quickly search textual information, ChemDB has been updated with an annotation parser and indexer implemented with Lucene (http://lucene.apache.org), which accommodates the special syntax often found in the chemistry domain, such as SMILES strings. This module provides full-text indexing and sub-second searching capabilities over the vendor annotations, comparable to those of a typical web-based document search engine. With this tool, users can retrieve full chemical structure records given only the chemical's name, CAS number or other identifier. Isoniazid and rifampicin and related information are easily found.

This is an especially convenient method for finding chemicals as they are generally identified by common, non-systematic, names that can only be indexed through an electronic knowledge repository, such as the corpus of vendor annotations. Systematic chemical naming protocols do exist based on the IUPAC standards (Skonieczny, 2006), and we do use OpenEye's Lexichem module (http://www.eyesopen.com) to generate systematic names. However, these are less useful for name lookup as users rarely search for '2-amino-3-phenyl-propanoic acid', instead of 'phenylalanine'. Moreover, in some cases, such as rifampicin, a large macrocyclic structure, systematic names cannot be generated reliably. ChemDB's full-text indexing of the vendor-supplied common names addresses these issues. Systematic names in ChemDB can nevertheless be useful when searching for multiple substring keywords. For instance, a keyword query such as 'amino phenyl propanoic' retrieves all chemicals whose name contains all of those keywords, effectively acting as a multiple functional group search.

As an additional convenience to users, the annotation search includes a 'fuzzy' search option. If the user is uncertain on how to spell complicated or ambiguous chemical names, or if the terms themselves have multiple spellings, the user can still submit a guess. The fuzzy search option matches annotations within a fixed edit distance from the query term. For example, searching for 'acetosalysilic acid' yields no results for most text-based search methods since no chemical exists by that name. With the fuzzy search option turned on, such a query returns several results, including the intended match for the structure of aspirin, with a correct name annotation of 'acetylsalicylic acid' that ChemDB rapidly identifies as a similar matching entry.

3.3 Searching virtual chemical space

Andrimid (Fig. 1) is a natural product with antibiotic activity capable of targeting fatty acid biosynthesis (Pohlmann *et al*., 2005). Thus andrimid, or andrimid-analogs, may be worth investigating as new potential tuberculosis drug leads. However, neither andrimid nor any analogs are found in ChemDB, suggesting that these compounds are not readily available commercially.

With ChemDB cataloging most commercially available chemicals, its contents represent a large fraction of 'real' chemical space. In this new release, ChemDB's exploratory capabilities are further extended to enable dynamic searches of 'virtual' chemical space, consisting of chemicals that are not already cataloged in the database, but which should be accessible by applying one or more synthetic chemical reactions to the commercially available building blocks contained in ChemDB.

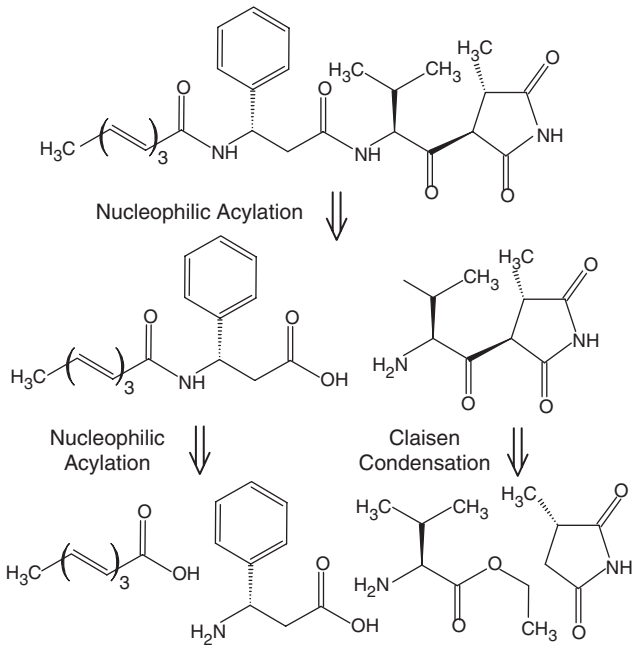

Fig. 1. Retrosynthetic deconstruction of andrimid, a natural product, suggested by ChemDB's virtual chemical space search.

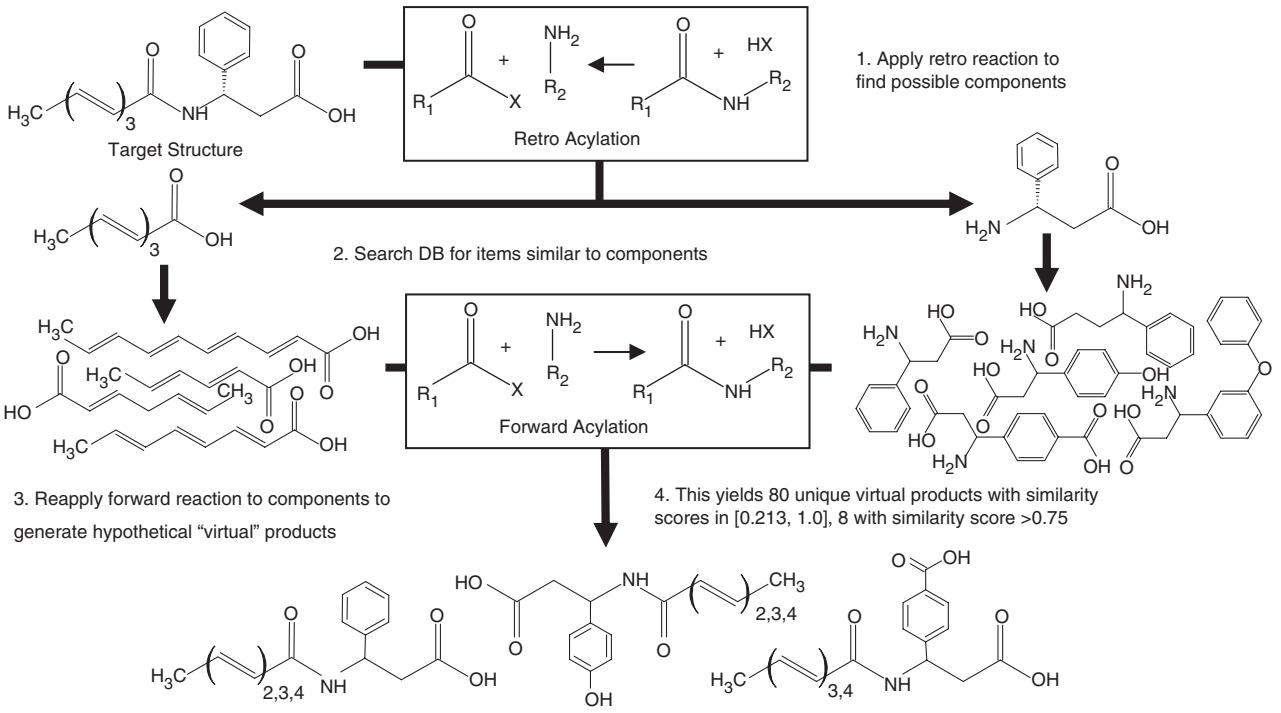

Fig. 2. Virtual chemical search applied to combinatorial library design. The overall goal, for instance within a drug discovery pipeline, is to find molecules similar to the target structure (a component of the natural product andrimid). A direct similarity search through the 'real' chemical space of ChemDB yields no results with Tanimoto similarity greater than 0.6. Rather than abandoning this lead, a virtual chemical search is run, using appropriate retro-reactions (retro Nucleophilic Acylation) to deconstruct the target into putative synthetic building blocks. Real starting materials are found by similarity search through ChemDB. The starting materials are then reassembled combinatorially using the corresponding forward reaction. Among the resulting 'virtual' products, 8 have >0.75 similarity to the target, including one exact match.

Focusing on a collection of reaction transformation models representing simple, but powerful, combinatorial reaction patterns (Corey *et al*., 1985), we have developed a collection of reagent models based upon the SMIRKS transformation language (James *et al*., 2005), with additional rule structures to handle forward and retro reaction processing and to help identify feasible reactions. One approach to searching virtual chemical space would be to simply apply all of these reaction models against every chemical combination in the database, producing the first iteration of 'Virtual ChemDB' and then apply the existing search algorithms to this dataset. However, with a database already containing ~5M compounds, the first iteration alone would yield on the order of 10^{13} virtual chemicals, an intractable amount to store and search for the foreseeable future.

Rather than generating the combinatorial explosion of all possible virtual products, ChemDB supports a targeted approach, exemplified in Figure 1 by the deconstruction of andrimid. Virtual searches begin with a user-supplied query molecule to find a synthesis pathway or generate a library of homologs. Given the target molecule, the system automatically finds 'retro' reaction models that can be applied to iteratively deconstruct the target into simpler precursors. Figure 2 illustrates an extension of this method where standard similarity searches through the database of real chemicals are done on the precursors and the most similar results are reapplied through the forward reaction model to yield hypothetical 'virtual' products. By construction, these products can be expected to bear some similarity to the original query molecule. This approach is related to retrosynthesis problem-solving (Todd, 2004), except that here we look for *any* molecule with *similarity* to the precursor molecules, whereas in retrosynthesis one is interested only in *exact* matches. In this manner, rather than suggesting only an exact synthetic pathway for the query molecule, the system can suggest a collection of building block precursors and reactions to build a combinatorial library of hypothetical products occupying a region of virtual chemical space around the query molecule.

ACKNOWLEDGEMENTS

Work supported by an NIH Biomedical Informatics Training grant (LM-07443-01) and NSF grants EIA-0321390 and 0513376 to P.B. We acknowledge the OpenBabel project, OpenEye Scientific Software, Peter Ertl of Novartis (JME Editor) and the JMol project for academic software licenses. We thank the laboratory of Dr Tsai for the tuberculosis drug discovery collaboration.

Conflict of Interest: none declared.

REFERENCES

Azencott,C.-A. *et al*. (2007) One- to four-dimensional kernels for small molecules and predictive regression of physical, chemical, and biological properties. *J. Chem. Inf. Model.*, **47**, 965–974.

Chen,J. *et al*. (2005) ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics*, **21**, 4133–4139.

Corey,E.J. *et al*. (1985) Computer-assisted analysis in organic synthesis. *Science*, **228**, 408–418.

Irwin,J.J. and Shoichet,B.K. (2005) ZINC – a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, **45**, 177–182.

James,C.A. *et al*. (2005). Daylight theory manual. http://www.daylight.com/dayhtml/doc/theory/

Lin,T.-W. *et al*. (2006) Structure-based inhibitor design of AccD5, an essential acyl-CoA carboxylase carboxyltransferase domain of Mycobacterium tuberculosis. *Proc. Natl Acad. Sci. USA*, **103**, 3072–3077.

Pohlmann,J. *et al*. (2005) Pyrrolidinedione derivatives as antibacterial agents with a novel mode of action. *Bioorg. Med. Chem. Lett.*, **15**, 1189–1192.

Skonieczny,S. (2006) The IUPAC rules for naming organic molecules. *J. Chem. Educ.*, **83**, 1633.

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

Swamidass,S.J. and Baldi,P. (2007a) Bounds and algorithms for fast exact searches of chemical fingerprints in linear and sub-linear time. *J. Chem. Inf. Model.*, **47**, 302–317.

Swamidass,S.J. and Baldi,P. (2007b) Mathematical correction for fingerprint similarity measures to improve chemical retrieval. *J. Chem. Inf. Model.*, **47**, 952–964.

Todd,M.H. (2004) Computer-aided organic synthesis. *Chem. Soc. Rev.*, **34**, 247–266.

Voigt,J.H. *et al*. (2001) Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.*, **41**, 702–712.