

CKB – The Compound Knowledge Base: A Text Based Chemical Search System

Matthew J. Walker,* Richard D. Hull,[†] and Suresh B. Singh

Molecular Systems Department, Merck Research Laboratories, P.O. Box 2000, Rahway, New Jersey 07065

Received April 25, 2002

The Compound Knowledge Base (CKB) was developed as a means of locating structures and additional relevant information from a given known structural identifier. Any of Chemical Abstracts Service Registry Number, company code (code number the producing company refers to the chemical entity internally), generic name (trivial or class name), or trade name (name under which the compound is marketed) can be provided as a query. CKB will provide the remaining available information as well as the corresponding structure for any matching compound in the database. The interface to the Compound Knowledge Base is Internet/World Wide Web-based, using Netscape Navigator and the ChemDraw Pro Plugin, which allows Merck scientists quick and easy access to the database from their desktop. The design and implementation of the database and the search interface are herein detailed.

INTRODUCTION

One rate-limiting step in the beginning of a computational study is the acquisition of the coordinates or connection tables for molecules of interest. For studies in which the number of molecules is small it is not onerous to look up structures in the literature and draw them from scratch, assuming one is able to locate the structures, which is often difficult. However, once the number of molecules goes from the tens to the thousands, drawing the molecules becomes impractical. There are many commercial and proprietary databases of chemicals available; however, often in the beginning of a study one has only the trivial or trade name of a compound. While most commercial database search engines are competent at searching based on a key or on structure, few allow for searches by textual term or pattern. Another issue is that there is no easy way to simultaneously search each of the available databases. Instead, one must repeat searches for each database in turn. This is not only tedious, but it may lead to redundant information that needs consolidation. Additionally, the name one has for a molecule might be outdated, retired, or not maintained by the database. For example, the original Pfizer company code for Celebrex is CP-X. A chemist or biologist reading a paper that mentions this compound might not realize that this company code represents Celebrex, and some databases will not associate CP-X with Celebrex. To address these limitations, we devised the Compound Knowledge Base (CKB),¹ a simple text database of chemical name data, with a simple, flexible Web-based method for query.

DESIGN AND IMPLEMENTATION

The first decision to be made in designing a database of this type is what to use as a primary key. Initially we considered using the Chemical Abstracts Service (CAS)

Registry Number (RN)² as a primary key for the text data. This has several limitations however, as few molecules in the commercial databases actually have CAS RNs associated with them. In addition the CAS RNs in the commercial databases are sometimes unreliable, in that the reported CAS RN is misassigned (that for the parent form instead of a salt) or are simply not valid. A more general key for the data is one calculated from the connection table as this guarantees the retrieved structure and the key match.

Topological Hash. One such key is the topological hash (topohash)³ for the molecule, a proprietary descriptor calculated from the connectivity of the molecule. For a molecule with m atoms, an $(m \times m)$ matrix X is constructed so that diagonal element $x(i,i)$ has the value

$$\begin{aligned} &1 * (\text{the atomic number for atom } i) + \\ &0.1 * (\text{number of heavy atom neighbors for atom } i) + \\ &\quad * (\text{the number of pi electrons for atom } i) + \\ &0.001 * (\text{match state, always 0 in this application}) + \\ &0.0005 * (\text{chirality, 1 for R and } -1 \text{ for S}) \end{aligned}$$

and the off-diagonal element $x(i,j)$ has the value: c/dist where dist is the through-bond distance between atom i and atom j and c is a user-supplied constant (0.4 in our case). The hash string is of the form m_E1_Em where $E1$ is the smallest eigenvalue of X and Em is the largest eigenvalue. The topological hash was calculated for each member of each database of interest and stored with the connection table and other data. With any hashing function, collisions are possible, but we expect them to be infrequent with this algorithm.⁴ When collisions do occur, they are not detected or treated, and the records are just assumed to be the same entity.

Database Interface. A Perl⁵ module was designed that mines a flat text representation of databases using Perl regular expressions⁶ to find data of interest, specifically the following: CAS RNs, company codes, generic names, trade names, and the registration code for each molecule, in addition to

* Corresponding author phone: (732)594-8575; fax: (732)594-4224; e-mail: Matthew_Walker@merck.com.

[†] Present address: Hull Consulting, Inc., 2646 Windsorgate Lane, Orlando, FL 32828.

the newly calculated topohash. These retrieved data were then “compressed” into a single record, which consists of the topohash, CAS RNs, generic names, trade names, and company codes separated by a colon and with unique data in each field separated by a semicolon. As a part of the compression, dashes, underscores, commas, and hashes are removed⁷ from company codes. All other terms are processed by compressing whitespace to a single space and removing terminal periods. All fields are made lowercase. The topohash is used as the first field and as the primary key for the database. As such, records for which no topohash could be calculated are excluded. The second field is the name of the source database from which the data was culled and the registration code of the molecule in that database. The remaining fields are as described above. The resulting string is as follows:

```
topohash:source=id;...:cas;...:generic;...:trade;...:code;...
```

These records are stored in a Berkeley DB⁸ format database using the Perl DB_File⁹ interface. The Perl DB_File interface allows the underlying dbm database to be “tied” via the Perl tie function to a Perl associative array (hash). The database then can simply be addressed by using standard Perl functions that work on hashes. Writing new database entries is as simple as a variable assignment: `$DATA{$key} = $value`.

This “primary” database described above is then used to produce a secondary, inverted database. Each record of the primary database is split into terms (i.e. each individual CAS RN, generic name, trade name, company code). Each term is then stored as a key of a record in the new secondary database. The value for this record is the key from the primary database record from which the term came. In the case that the term occurs in more than one record in the primary database, the value will be the keys from all matching records concatenated with semicolons. The primary and secondary databases together are henceforth cumulatively referred to as the CKB database.

The process of updating or adding a new data source to the database is the same as that of creating it. A new data source is analyzed to find the fields of interest. Appropriate modifications are made to the process function to mine these fields from the database, and the resulting data are compressed into the primary database. The inverted database must then be recreated from scratch in order to pick up new terms.

The current version (v2.2) of the CKB database consists of the fusion of the Merck Index 2000 (Merck in-house special edition), MDL Comprehensive Medicinal Chemistry,¹⁰ MDL Drug Data Report,¹¹ National Cancer Institute Developmental Therapeutics Program,¹² and Standard Drug File¹³ databases covering approximately 250 000 records.

Using CKB. The primary means of accessing the CKB is through a Web interface. Choosing a Web interface for CKB was easy as Web browsers are nearly ubiquitous and programming hypertext markup language (HTML) is simple. Deciding the means for display of structures was more difficult. While current browsers can render structures as images, we wanted to provide users with the ability to copy the structures resulting from searches into other applications. The ChemDraw Pro Plugin (CDP)¹⁴ was chosen for the

Figure 1. CKB query form with zocor as a query term.

display of structures, as it provides the complete tools of the ChemDraw program embedded with a Web page.

Search Form. The search interface provides a very simple HTML form (Figure 1) which allows the user to provide either a term or pattern search. The term search simply looks for exact matches in the inverted database, and hits are used to look up full records in the main database. With the pattern search, one can provide a Perl regular expression⁶ as a query which is used to search each full record in the primary database, one record at a time. Regular expression languages such as the one implemented in Perl are used to specify patterns, often with wildcards, for searching data. Exact and partial matches can be searched for in this manner at the expense of speed.

RESULTS

Simple Term Search. The simplest and most common search method is a term search. This kind of search is useful if the chemist knows a name or names for the compound of interest and wants to quickly retrieve any additional information. A search on the term “zocor” reports (in less than 6 s) a single hit with CAS RN 79902-63-9 and trade names such as Liponorm (Figure 2). The first column is a representation of the structure of the molecule as retrieved from the structure database in MDL MOL¹⁵ format and displayed by the CDP. The second “Seeks” column contains links to proprietary internal search engines. The third column is a list of the CAS RNs associated with this record. There can be more than one CAS RN for reasons described above. The fourth “Source and ID” column contains a list of the databases from which the data for this record was culled, along with a link allowing the original record to be retrieved. The setting specified by the “Source and ID link sends:” drop down in the search form (Figure 1) determines the format of the data returned. The final columns are the generic names, trade names, and company codes, respectively. The Merck Index monograph number is listed as a company code with the prefix mono, mono8686 in this case. Because the terms are processed into the inverted database with little modification, search terms must be exact. A query on “dextromethorphan” will find no hits because the only mention of the term “dextromethorphan” in the database is as “dextromethorphan hydrobromide”. Similarly, if dashes were present in the record they were retained in the term. Searching for “cocaine hydrochloride” will get no hits while “cocaine-hydrochloride” will.

Pattern Search. The pattern search allows the chemist to locate a chemical when the complete name is not known. If, for instance, the chemist is uncertain of the spelling of a trade name, so the unspecified parts can be treated with wildcards, such as the query: `g.*fen.*n` in which the “.” can be any character and the “*” indicates 0 or more of the previous character. This pattern would match `gfenn` as well

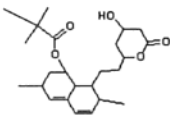
Structure	Seeks	CAS #	Source and ID	Generic Names	Trade Names	Company Codes
	<u>CS</u> - <u>JS</u>	79902-63-9	cmc = 6248 mddr = 122234 mddr = 165929 mindex2000 = 08686 mindex = 7134 sdf = SYNVINOLI	simvastatin	denan liponorm lodalas simvastatin sinvacor sivastatin sivastin synvinolin zocor zocord	mk733 mono08686 mono8686

Figure 2. Results of CKB term search on "zocor".

as the difficult to spell expectorant guaifenesin. Another use for the pattern search is to retrieve classes of similarly named compounds, such as all of the statins. A search with the query *.statin retrieves over 300 molecules. This search method can be very flexible but is CPU intensive with an average query taking about 60 s in the Web interface, as opposed to about 6 s for a term search.

Future Directions. The CKB database is simple to update and extend. To add another database a data processor routine must be written to extract the relevant information from the new database and the database reconstructed, a process, which currently takes about 30 min. As additional databases become available they can be added to the CKB. Another direction for future development is to improve the parsing of the terms into the secondary database by removing very common words such as "the" and by better regularizing the data.

DISCUSSION AND CONCLUSIONS

Expedient access to data is a very important issue for modeling in the pharmaceutical industry. Because of the time constraints of medicinal chemistry projects, modelers cannot afford to invest time tracking down known information. While there are several chemical database systems available that allow searching of text data such as ISIS¹⁶ or Thor,¹⁷ there are none that allow quick and simultaneous searches of several commercial databases as well as proprietary in-house data sources in such a flexible manner. The CKB has an advantage over a commercial product such as Thor in that it uses standard tools and interfaces which are available on a wide variety of platforms. Almost all flavors of the UNIX operating system support Perl, BerkeleyDB, and Web servers such as Netscape or Apache. Solutions using commercial products are limited to architectures and operating systems supported by the vendor.

The CKB database provides a simple, easy to use, Web-based interface which allows the chemist to search for known chemical structures across a number of databases. With a single piece of information, such as a company code or a CAS RN, many additional data including the structure of many compounds can be quickly retrieved.

ACKNOWLEDGMENT

We thank Robert Sheridan for useful discussions on the use of topological hashes for primary keys.

REFERENCES AND NOTES

- (1) The term "Knowledge" in the name was given in the somewhat whimsical sense of "given what I know about this molecule, give me the rest...". In the formal sense the more correct term would be "Information". However, "Compound Information Base" name is already used for another project within Merck.
- (2) Chemical Abstract Service, <http://www.cas.org>.
- (3) (a) Sheridan, R. P. The Most Common Chemical Replacements in Drug-Like Compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103–108. (b) Burden, F. R. A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix. *QSAR* **1997**, *16*, 309–314. (c) Burden, F. R. Molecular Identification Number for Substructure Searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225–227.
- (4) Note ref 3c above discusses the occurrence of collisions, with no collisions in the 18 029 alkanes with 16 or fewer carbons (with lowest two eigenvalues to eight places) and two collisions in the 1560 alkenes of 11 carbons and less. We found collisions to occur with only 0.4% of the molecules in the Merck Corporate database.
- (5) (a) Wall, L. Practical Extraction and Report Language, version 5.005_03; <http://www.perl.com>. (b) Wall, L.; Christiansen, T.; Schwartz, R. L. *Programming Perl*, 2nd ed.; O'Reilly & Associates, Inc.: 1996.
- (6) An excellent general description of regular expressions can be found in the following: Friedl, J. E. F. *Mastering Regular Expressions*; O'Reilly & Associates, Inc.: 1997.
- (7) Perl regular expression: `s/[-_#]//`.
- (8) Berkeley DB, version 2.77, Sleepycat Software, <http://www.sleepycat.com>.
- (9) DB_File.pm, version 1.71, Marquess, P., <http://cpan.perl.com>.
- (10) Comprehensive Medicinal Chemistry Database version 1999.1, <http://www.mdl.com/products/cmc.html>, MDL Information Systems, Inc., <http://www.mdl.com>.
- (11) MDL Drug Data Report version 1999.2, <http://www.mdl.com/products/mddr.html>, MDL Information Systems, Inc., <http://www.mdl.com>.
- (12) NCI Developmental Therapeutics Program Database, version 1994.1a, MDL Information Systems, Inc., <http://www.mdl.com>.
- (13) Standard Drug File, MACCS version ca. 1997, (discontinued product), MDL Information Systems, Inc., <http://www.mdl.com>.
- (14) ChemDraw Pro Plugin, version 5.0, CambridgeSoft Corporation., <http://www.cambridgesoft.com>.
- (15) MOLfile, http://www.mdl.com/downloads/ctfile/ctfile_subs.html, MDL Information Systems, Inc., <http://www.mdl.com>.
- (16) Integrated Scientific Information System, <http://www.mdl.com/products/isis.html>, MDL Information Systems, Inc., <http://www.mdl.com>.
- (17) Thor, Daylight CIS, Inc., <http://www.daylight.com>.

CI0255329