WebChemDB: An Integrated Chemical Database Retrieval System

Bo Kyeng Hou1*, Eun-joung Moon2, Sung Chul Moon2 and HaeJin Kim2

1Korean Bioinformation Center, KRIBB, Daejeon 305-806, 2Ensoltek, Daejeon 305-510, Korea

Abstract

WebChemDB is an integrated chemical database retrieval system that provides access to over 8 million publicly available chemical structures, including related information on their biological activities and direct links to other public chemical resources, such as PubChem, ChEBI, and DrugBank. The data are publicly available over the web, using two-dimensional (2D) and three-dimensional (3D) structure retrieval systems with various filters and molecular descriptors. The web services API also provides researchers with functionalities to programmatically manipulate, search, and analyze the data.


Keywords: Chemical database, biological activities, structure retrieval system, molecular descriptors

Introduction

Small molecules can be used as building blocks for combinatorial chemical synthesis; as molecular probes for analyzing biological systems in systems biology; and for the screening, design, and discovery of useful drug compounds (Chen et al., 2005; Ahn, 2007; Kang et al., 2009). Tying together many disparate sources of chemical and life sciences data into an integrated database is one of the main issues in the bioinformatics community. Large public chemical databases, such as PubChem (Wang et al., 2009) and DrugBank (Wishart et al., 2008), provide chemical structures and associated biological information, focusing on small organic molecules that have potential use in drug development with biomedical research.

PubChem is an increasingly popular, free-access, online molecular database that is operated by the National Center for Biotechnology Information (NCBI). But, many of the kinds of information that biologists find most interesting (links to primary literature; characterization data in the form of spectra, solubilities, melting/boiling points, etc.) do not appear in PubChem.

Although the existing databases, in their current form, consist mainly of a catalog of biologically relevant molecules, increasing the level of crosslinking to other biological databases could result in a much more useful service. As chemical databases that contain intersecting information continue to proliferate, such crosslinking is likely to increase in importance (http://zusammen.metamodel-database.com).

This paper, therefore, presents an integrated chemical database retrieval system for searching, visualizing, and analyzing chemical structures with associated biological information, including precalculated values for molecular properties (e.g., 3D coordinates, molecular weight, polar surface area, hydrogen bond donors/acceptors, rotatable bonds, XLogP, etc.). The system also offers a focused subset of calculated properties (e.g., drug-like, lead-like, etc.) for molecules that are filtered by physical properties. As such, it may be useful to those who perform docking experiments or build focused chemical databases.

Data and processing

The data were collected from and cross-referenced to public chemical data resources, such as PubChem (Wang et al., 2009), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2000), the Chemical Entities of Biological Interest (ChEBI) (Degtyarenko et al., 2008), NMRShiftDB (Steinbeck et al., 2004), Distributed Structure-Searchable Toxicity (DSSTox) (Richard et al., 2002), and DrugBank (Wishart et al., 2008).

The public chemical data are available in various data formats (e.g., MOL, SDF, SMILES, etc.). Therefore, our parsing tool was used to parse the public chemical data to extract data according to field name mapping table, which were then dynamically uploaded to the database. The database schema consisted of chemical identification information and associated data, ranging from molecular descriptors to pathway data, spectroscopic data, and toxicological data. The molecular descriptors (3D coordinates, hydrogen-bond donors, hydrogen-bond acceptors, octanol/water partition coefficient log P, etc.) that were not available in the imported data were calcu-
Table 1. Subsets of WebChemDB filtered by physical properties

<table>
<thead>
<tr>
<th>Filter subsets</th>
<th>Hit compounds</th>
<th>Selection criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug-like (Lipinski, 2000)</td>
<td>4,652,437</td>
<td>150 &lt; Molecular weight &lt;=500 and Hydrogen-bond acceptors &lt;=10 and xLogP &lt;=5 and Rotatable bonds &lt;= 8 and Polar surface area &lt;=150</td>
</tr>
<tr>
<td>Newton-hit-like (Irwin et al., 2005)</td>
<td>531,930</td>
<td>200 &lt; Molecular weight &lt;=350 and 1 &lt; xLogP &lt;=3</td>
</tr>
<tr>
<td>Fragment-like (Carr et al., 2005)</td>
<td>201,116</td>
<td>150 &lt; =Molecular weight &lt;=250 and Hydrogen-bond acceptors &lt;=4 and Rotatable bonds &lt;= 3 and xLogP &lt;=3 and Polar surface area &lt;=200</td>
</tr>
<tr>
<td>Greasy-leads (Irwin et al., 2005)</td>
<td>431,162</td>
<td>150 &lt;= Molecular weight &lt;=350 and 2 &lt; xLogP &lt;=6</td>
</tr>
<tr>
<td>Lead-like (Teague, 1999)</td>
<td>1,788,861</td>
<td>150 &lt; =Molecular weight &lt;=350 and Hydrogen-bond acceptors &lt;=6 and Hydrogen-bond donors &lt;= 3 and xLogP &lt;=4 and Rotatable bonds &lt;= 3</td>
</tr>
<tr>
<td>Big-n-greasy (Irwin et al., 2005)</td>
<td>1,270,196</td>
<td>300 &lt; =Molecular weight &lt;=600 and 2 &lt; xLogP &lt;=6</td>
</tr>
</tbody>
</table>

Fig. 1. Screenshot of the webpage. The chemical structures shown in the sketcher window of the two-dimensional (2D) search can be searched in the three search types (Exact, Substructure, and Similarity). The similarity threshold is accomplished by setting the parameters of the similarity measure to 0% and 100% in the similarity search type. The search option has maximum hits and maximum time to restrict search spaces. The search is integrated with molecular descriptors as standard filters, like those shown to restrict the results by molecular weight, hydrogen-bond acceptor, hydrogen-bond donor, predicted XLogP, rotatable bonds, and polar surface area. The user can specify ranges for combinations of five molecular descriptors. The Descriptor button shows molecular descriptors of the query compound.

Query Tool and User Interface

The database can search over 8 million compounds through the structure retrieval systems, which provide capabilities for: (1) two-dimensional (2D) structure searches; (2) three-dimensional (3D) structure searches (Raymond et al., 2003); (3) text searches; and (4) filter searches. The three search types that are available on the 2D structure search page are as follows: (1) Substructure Search (the default): the search for all compounds that have that “substructure” and other atoms, (2) Similarity Search: the search for all compounds that contain structural features that are similar, based on the Tanimoto coefficient (Butina, 1999), (3) Exact Structure: the search for all compounds that are exact structure matches in the database. The text search retrieves all

lated by using the calculation modules that were available from CDK (Steinbeck et al., 2003). The detailed information on the database schema and records are available at http://biocommunity.kr/chemsearch/schema.pdf. The database was implemented using the leading open-source relational database MySQL (http://www.mysql.com).
compounds by using text-matching capabilities. The filter search finds useful drug compounds by using known filters, such as drug-like, newton-hit-like, fragment-like, greasy-leads, lead-like, and big-n-greasy, as shown in Table 1.

The proposed system uses MarvinSketch (http://www.chemaxon.com/product/msketch.html) for drawing molecules and queries, MarvinView (http://www.chemaxon.com/product/mview.html) for viewing single and multiple chemical structures, and CDK for calculating physicochemical properties (e.g., molecular weight, hydrogen-bond acceptor/donors, etc.) and filtering database search results. The two-dimensional (2D) structure search engine and structure database tables are made by using the JChem library (http://www.chemaxon.com/jchem/intro/index.html) from ChemAxon, Inc. The search engine will then retrieve all of the molecules in the database that contain the specified chemical structure. Data from associated property files can be displayed along with the structures or searched as an additional parameter. Web interfaces are delivered using the open-source Tomcat Web server (http://tomcat.apache.org). The web interface provides the user with a wide array of filters and threshold values to be tailored to different searches (Fig. 1). A user can either draw the compound to be searched for using MarvinSketch or enter its SMILES string, and choose the type of structure search.

**Fig. 2.** Results page of a 2D structure search. The search result can be exported through the ‘Export’ button. The ‘More’ button provides detailed information on the chemical structure.
Table 2. Web services API of WebChemDB

<table>
<thead>
<tr>
<th>Function name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fconv_2Dsdff_2_3Dsdff</td>
<td>Convert an SDF file for 2D molecules to an SDF file for 3D molecules (Csizmadia, 2000)</td>
</tr>
<tr>
<td>Similarity_3Dsearch_by_3Dsdff</td>
<td>Search compounds that have a similar three-dimensional structure to the query compound (Raymond et al., 2003)</td>
</tr>
<tr>
<td>Get_compound_by_kcid</td>
<td>Get a compound in SDF format by compound ID (kcid)</td>
</tr>
<tr>
<td>Get_substance_by_ksid</td>
<td>Get a substance in SDF format by substance ID (ksid)</td>
</tr>
<tr>
<td>Substructure_2Dsearch_by_2Dsdff</td>
<td>Search substructures by query compound (Csizmadia, 2000)</td>
</tr>
<tr>
<td>Similarity_2Dsearch_by_2Dsdff</td>
<td>Search similar compounds by chemical hashed binary fingerprint and Tanimoto coefficient (Butina, 1999)</td>
</tr>
<tr>
<td>Exact_2Dsearch_by_2Dsdff</td>
<td>Search structures that match the given compound exactly (Csizmadia, 2000)</td>
</tr>
<tr>
<td>Chemical_search_by_text</td>
<td>Search compounds by text</td>
</tr>
<tr>
<td>Get_descriptors_by_sdf</td>
<td>Calculate molecular descriptors of given compound</td>
</tr>
</tbody>
</table>

Conclusions

WebChemDB is an integrated chemical database that contains over 8 million small molecules, collected and crosslinked from public chemical resources. The data are publicly available on the worldwide web for download and for targeted searches using a variety of powerful search engines. The chemical data include predicted or experimentally determined physicochemical properties, such as 3D structures and predicted xLogP. Web services API can help users to programmatically manipulate, search, and analyze chemical structure data in WebChemDB. In the future, we will make a workflow extension package for workflow systems, such as Taverna (Hull et al., 2006) and KNIME (Michael et al., 2008), that enables users to visually create data flows, selectively execute some or all analysis steps, and investigate the results later through interactive views on data and models.

Acknowledgments

We thank Jimin Hwang for editing the manuscript. This project was supported by the Korean Ministry of Education, Science and Technology (MEST) under grant number 2009-0080141.

References


