APPLICATION NOTE Genomics & Informatics Vol. 7(4) 212-216, December 2009

WebChemDB: An Integrated Chemical Database Retrieval System

Bo Kyeng Hou[1]*, Eun-joung Moon[2], Sung Chul Moon[2] and HaeJin Kim[2]

[1]Korean Bioinformation Center, KRIBB, Daejeon 305-806, [2]Ensoltek, 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.

Availability: The database is accessible at http://bio-community.kr/chemsearch/index.jsp.

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.meta-molecular.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-

*Corresponding author: E-mail bkher71@kribb.re.kr
Tel +82-42-879-8526, Fax +82-42-879-8519
Accepted 22 November 2009

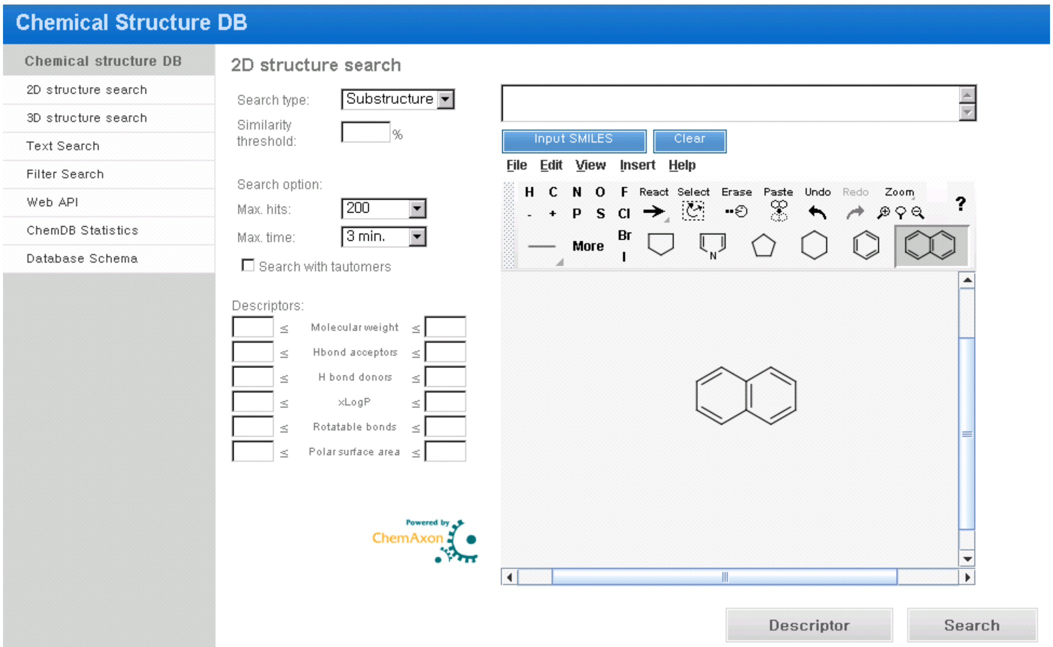
Table 1. Subsets of WebChemDB filtered by physical properties

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

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).

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

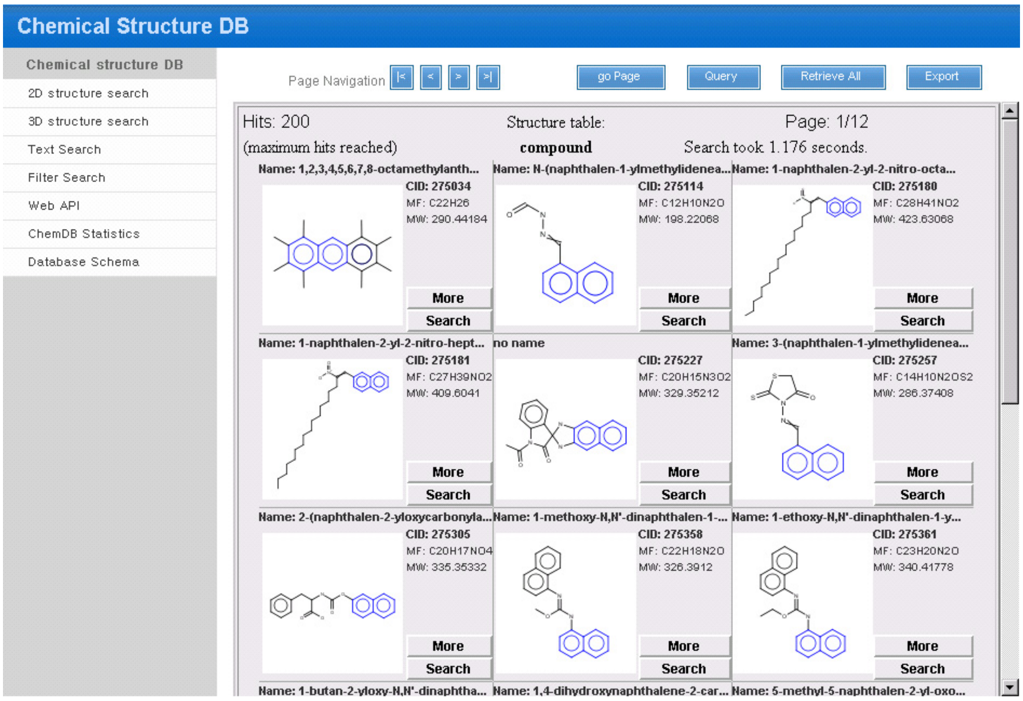


Search type: Substructure

Max. hits: 200

Max. time: 3 min.

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 donors, 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.



Hits: 200 (maximum hits reached)

Structure table: **compound**

Page: 1/12

Search took 1.176 seconds.

Name: 1,2,3,4,5,6,7,8-octamethylanth...
CID: 275034
MF: C22H26
MW: 290.44184

Name: N-(naphthalen-1-ylmethylidenea...
CID: 275114
MF: C12H10N2O
MW: 198.22068

Name: 1-naphthalen-2-yl-2-nitro-octa...
CID: 275180
MF: C28H41NO2
MW: 423.63068

Name: 1-naphthalen-2-yl-2-nitro-hept...
CID: 275181
MF: C27H39NO2
MW: 409.6041

no name
CID: 275227
MF: C20H15N3O2
MW: 329.35212

Name: 3-(naphthalen-1-ylmethylidenea...
CID: 275257
MF: C14H10N2OS2
MW: 286.37408

Name: 2-(naphthalen-2-yloxycarbonyla...
CID: 275305
MF: C20H17NO4
MW: 335.35332

Name: 1-methoxy-N,N'-dinaphthalen-1-...
CID: 275358
MF: C22H18N2O
MW: 326.3912

Name: 1-ethoxy-N,N'-dinaphthalen-1-y...
CID: 275361
MF: C23H20N2O
MW: 340.41778

Name: 1-butan-2-yloxy-N,N'-dinaphtha...

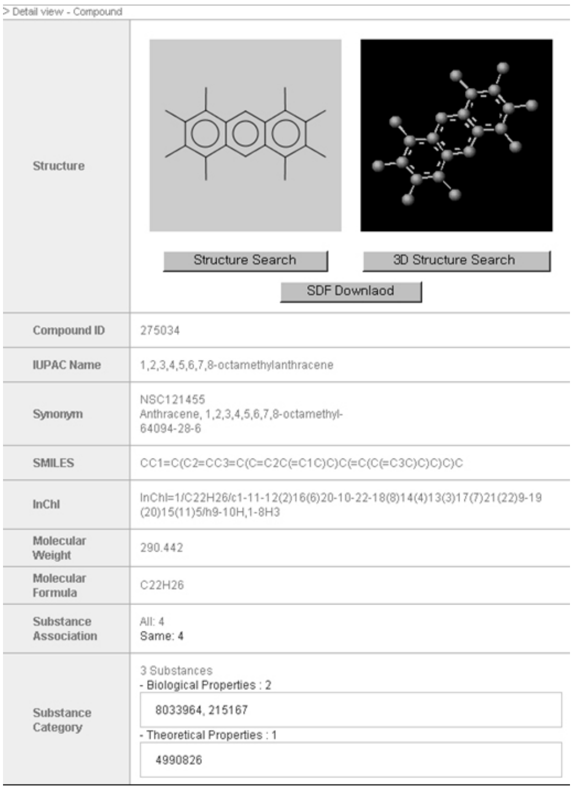
Name: 1,4-dihydroxynaphthalene-2-car...

Name: 5-methyl-5-naphthalen-2-yl-oxo...

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.

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



Structure	
Compound ID	275034
IUPAC Name	1,2,3,4,5,6,7,8-octamethylanthracene
Synonym	NSC121455 Anthracene, 1,2,3,4,5,6,7,8-octamethyl- 64094-28-6
SMILES	CC1=C(C2=CC3=C(C=C2C(=C1C)C)C(=C(C(=C3C)C)C)C)C
InChI	InChI=1/C22H26/c1-11-12(2)16(6)20-10-22-18(8)14(4)13(3)17(7)21(22)9-19(20)15(11)5/h9-10H,1-8H3
Molecular Weight	290.442
Molecular Formula	C22H26
Substance Association	All: 4 Same: 4
Substance Category	3 Substances - Biological Properties : 2 8033964, 215167 - Theoretical Properties : 1 4990826

Fig. 3. Detailed information for a chemical structure. Shown are the compound identification number, 3D structure, synonym, and crosslinks of biological property, etc.

Table 2. Web services API of WebChemDB

Function name	Description
Fconv_2Dsdf_2_3Dsdf	Convert an SDF file for 2D molecules to an SDF file for 3D molecules (Csizmadia, 2000)
Similarity_3Dsearch_by_3Dsdf	Search compounds that have a similar three-dimensional structure to the query compound (Raymond *et al.*, 2003)
Get_compound_by_kcid	Get a compound in SDF format by compound ID (kcid)
Get_substance_by_ksid	Get a substance in SDF format by substance ID (ksid)
Substructure_2Dsearch_by_2Dsdf	Search substructures by query compound (Csizmadia, 2000)
Similarity_2Dsearch_by_2Dsdf	Search similar compounds by chemical hashed binary fingerprint and Tanimoto coefficient (Butina, 1999)
Exact_2Dsearch_by_2Dsdf	Search structures that match the given compound exactly (Csizmadia, 2000)
Chemical_search_by_text	Search compounds by text
Get_descriptors_by_sdf	Calculate molecular descriptors of given compound

(similarity, substructure, or exact) to do and the similarity threshold to use. The structures of matching compounds are returned (Fig. 2). Clicking on a returned structure leads to detailed information on the corresponding compound (Fig. 3). Many of the application tools, scripts, and Web interfaces are written in Java (http://java.sun.com), SQL, and JSP (JavaServer Pages).

The web services API

In recent years, Internet technology for application-to-application communication, commonly referred to as web services, has expanded rapidly due to growing needs (Shuichi *et al.*, 2003). Web services enable users to develop software that accesses and manipulates massive amounts of data that are constantly refreshed. We have developed a web service API using SOAP (Simple Object Access Protocol) and WSDL (Web Service Definition Language) to programmatically manipulate, search, and analyze chemical structure data in WebChemDB (Table 2). The WSDL file, web service client library, and sample source codes are available at http://biocommunity.kr/chemsearch/api.jsp.

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

Ahn, C. (2007). Pharmacogenomics in Drug Discovery and Development. *Genomics & Informatics* 5, 41-45.

Butina, D. (1999). Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* 39, 747-750.

Carr, R.A., Congreve, M., Murray, C.W., and Rees, D.C. (2005). Fragment-based lead discovery: leads by design. *Drug Discov. Today* 10, 987-992.

Chen, J., Swamidass, S.J., Dou, Y., Bruand, J., and Baldi, P. (2005). ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics* 21, 4133-4139.

Csizmadia, F. (2000). JChem: Java applets and modules supporting chemical database handling from web browsers. *J. Chem. Inf. Comput. Sci.* 40, 323-324.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., and Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucl. Acids Res.* 36, 344-350.

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M.R., Li, P., and Oinn, T. (2006). Taverna: a tool for building and running workflows of services. *Nucl. Acids Res.* 1, 729-732.

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

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucl. Acids Res. 28, 27-30.

Kang, T.S., Woo, S.W., Park, H.J., Han, S.Y., Park, M.H., and Chung, M.W. (2009). The Korean Pharmacogenomic Database at NIFDS: 2008 Update. *Genomics & Informatics* 7, 163-167.

Lipinski, J. (2000). Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* 44, 235-249.

Michael, R.B., Nicolas, C., Fabian, D., Thomas, R.G., Tobias, K., Thorsten, M., Peter, O., Christoph, S., Kilian, T., and Bernd, W. (2008). KNIME: The Konstanz Information Miner. *Data Analysis, Machine Learning and Applications* 319-326.

Raymond, J. W., and Willett, P. (2003). Similarity searching in databases of flexible 3D structures using smoothed bounded distance matrices. *J. Chem. Inf. Comput. Sci.* 43, 908-916.

Richard, A.M., and Williams, C.R. (2002). Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. *Mutat. Res.* 499, 27-52.

Shuichi, K., Toshiaki, K., Yoko, S., and Minoru, K. (2003). KEGG API: A Web Service Using SOAP/WSDL to Access the KEGG System. *Genome Informatics* 14, 673-674.

Steinbeck, C., and Kuhn, S. (2004). NMRShiftDB -- compound identification and structure elucidation support through a free community-built web database. *Phytochemistry* 65, 2711-2717.

Steinbeck, C., Han, Y. Q., Kuhn, S., Horlacher, O., Luttmann, E., and Willighagen, E.L. (2003). The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* 43, 493-500.

Teague, S.J., Davis, A.M., Leeson, P.D., and Oprea, T. (1999). The Design of Leadlike Combinatorial Libraries. Angew. *Chem. Int. Ed. Engl.* 38, 3743-3748.

Voigt, J.H., Bienfait, B., Wang, S., and Nicklaus, M.C. (2001). Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.* 41, 702-712.

Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J., and Bryant, S.H. (2009). PubChem: a public information system for analyzing bioactivities of small molecules. *Nucl. Acids Res.* 37, 623-633.

Weininger, D., Weininger, A., and Weininger, L.J. (1989). SMILES. 2. Algorithm for generation of uniques SMILES notation. *J. Chem. Inf. Comput. Sci.* 29, 97-101.

Wishart, D.S. (2008). DrugBank and its relevance to pharmacogenomics. *Pharmacogenomics.* 9, 1155-1162.