ISSN: 0975-766X

Available Online through
www.ijptonline.com

Research Article

CHEMSPIDER: A MAGIC BULLET OF CHEMICAL DATABASE FOR CHEMISTS

Palakben K. Parikh, Julee P. Soni, Kaumil N. Modi, Urviben Y. Patel Ravi N. Patel and Prof. Dr. Dhrubo Jyoti Sen*

Department of Pharmaceutical Chemistry, Shri Sarvajanik Pharmacy College, Gujarat Technological University, Arvind Baug, Mehsana-384001, Gujarat, India, Phone: 02762-247711, Fax: 02762-247712,

Email: *dhrubosen69@yahoo.com*

Received on 31-07-2010 ***Accepted on 18-08-2010***

Abstract:

Chemspider is a chemical database (Website: www.ChemSpider.com). The system was first launched in March 2007 in a beta release form and transitioned to release in March 2008. ChemSpider has expanded the generic support of a chemistry database to include support of the Wikipedia chemical structure collection via their WiChempedia implementation. ChemSpider was acquired by the Royal Society of Chemistry in May, 2009. Headquarters at Raleigh, North Carolina February, 2007. This is a very interesting database in the field of chemical science.

Keywords: Database, Wikipedia, Lipidomics, Kyoto Encyclopedia of Genes and Genomes, Xpharm

Database

The database contains more than 20 million unique molecules from the following sources:

- **A-L**: EPA DSSTox (Distributed Structure-Searchable Toxicity), U.S. Food and Drug Administration (FDA), Human Metabolome Database, Journal of Heterocyclic Chemistry, KEGG, KUMGM, LeadScope, LipidMAPS
- **M-N**: Marinlit, MDPI, MICAD, MLSMR, MMDB, MOLI, MTDP, Nanogen, Nature Chemical Biology, NCGC, NIAID, NIH (National Institute of Health)/NLM (National Library of Medicine), NINDS Approved Drug Screening Program (National Institute of Neurological Disorders and Stroke), NIST, NIST Chemistry WebBook, NMMLSC, NMRShiftDB.

- **P-S**: PANACHE, PCMD, PDSP, Peptides, Prous Science Drugs of the Future, QSAR, R&D Chemicals, San Diego Center for Chemical Genomics, SGCOxCompounds, SGCStoCompounds, SMID, Specs, Structural Genomics Consortium, SureChem, Synthon-Lab
- **T-Z**: Thomson Pharma, Total TOSLab Building-Blocks, UM-BBD, UPCMLD, UsefulChem, Web of Science, xPharm, ZINC

The database can be updated with user contributions including chemical structure deposition, spectra deposition and user curation.

Lipidomics may be defined as the large-scale study of pathways and networks of cellular lipids in biological systems[1]. The word "lipidome" is used to describe the complete lipid profile within a cell, tissue or organism and is a subset of the "metabolome" which also includes the three other major classes of biological molecules: proteins/amino-acids, sugars and nucleic acids. Lipidomics is a relatively recent research field that has been driven by rapid advances in technologies such as mass spectrometry (MS), nuclear magnetic resonance (NMR) spectroscopy, fluorescence spectroscopy, dual polarisation interferometry and computational methods, coupled with the recognition of the role of lipids in many metabolic diseases such as obesity, atherosclerosis, stroke, hypertension and diabetes. This rapidly expanding field complements the huge progress made in genomics and proteomics, all of which constitute the family of systems biology. Lipidomics research involves the identification and quantitation of the thousands of cellular lipid molecular species and their interactions with other lipids, proteins, and other metabolites. Investigators in lipidomics examine the structures, functions, interactions, and dynamics of cellular lipids and the changes that occur during perturbation of the system. Han and Gross first defined the field of lipidomics through integrating the specific chemical properties inherent in lipid molecular species with a comprehensive mass spectrometric approach[2]. Although lipidomics is under the umbrella of the

more general field of "metabolomics", lipidomics is itself a distinct discipline due to the uniqueness and functional specificity of lipids relative to other metabolites.

In lipidomic research, a vast amount of information quantitatively describing the spatial and temporal alterations in the content and composition of different lipid molecular species is accrued after perturbation of a cell through changes in its physiological or pathological state. Information obtained from these studies facilitates mechanistic insights into changes in cellular function. Therefore, lipidomic studies play an essential role in defining the biochemical mechanisms of lipid-related disease processes through identifying alterations in cellular lipid metabolism, trafficking and homeostasis. The growing attention on lipid research is also seen from the initiatives underway of the LIPID Metabolites And Pathways Strategy (LIPID MAPS Consortium) and The European Lipidomics Initiative (ELIfe).

Rosetta Genomics Ltd. (NASDAQ: ROSG) is a molecular diagnostics company with offices in Israel and the United States that uses micro-ribonucleic acid (microRNA) biomarkers to develop diagnostic tests designed to differentiate between various types of cancer. The company expects the first three tests based on its technology to be submitted for regulatory approval in 2008. The diagnostic tests will differentiate between squamous and non-squamous non-small cell lung cancer (NSCLC); differentiate between adenocarcinoma and mesothelioma; and seek to identify the origin of tumors in patients representing cancer of unknown primary (CUP)[3]. Using a single microRNA, the highly sensitive, highly specific test for squamous and non-squamous lung cancer has passed the prevalidation phase and has been submitted for approval to the New York State Department of Health Clinical Laboratory Evaluation Program in April 2008.

In April 2008, *Nature Biotechnology* published a study by Rosetta Genomics' scientists whose findings demonstrate microRNAs' significant potential to act as effective biomarkers that may be applied in a diagnostic test designed to identify the primary tumor site in patients CUP. In addition to its diagnostic programs, Rosetta Genomics is collaborating with Isis Pharmaceuticals to develop a microRNA-based therapy for Hepatocellular carcinoma (HCC), a form of liver cancer.

KEGG (**Kyoto Encyclopedia of Genes and Genomes**) is a collection of online databases dealing with genomes, enzymatic pathways, and biological chemicals. The PATHWAY database records networks of molecular interactions in the cells, and variants of them specific to particular organisms[4].

Introduction

The KEGG, the Kyoto Encyclopedia of Genes and Genomes, was initiated by the Japanese human genome programme in 1995. According to the developers they consider KEGG to be a "computer representation" of the biological system. The KEGG database can be utilized for modeling and simulation, browsing and retrieval of data. It is a part of the systems biology approach.

KEGG maintains five main databases[4]:

- KEGG Atlas
- KEGG Pathway
- KEGG Genes
- KEGG Ligand
- KEGG BRITE

Databases

KEGG connects known information on molecular interaction networks, such as pathways and complexes (this is the Pathway Database), information about genes and proteins generated by genome projects (including the gene database) and information about biochemical compounds and reactions (including compound and reaction databases). These databases are different networks, known as the protein network, and the chemical universe respectively. There are efforts in progress to add to the knowledge of KEGG, including information regarding ortholog clusters in the KO (KEGG Orthology) database[5].

KEGG Pathways:

- Metabolism
- Genetic Information Processing
- Environmental Information Processing
- Cellular Processes
- Human Diseases
- Drug development

Ligand Database:

- Compound
- Drug
- Glycan
- Reaction
- RPAIR (Reactant pair alignments)
- Enzyme

Xpharm: MDL Information Systems, a provider of R&D informatics offerings for the life sciences and chemicals industries and acquired by Symyx Technologies, Inc. in 2007, was launched as a computer-aided drug design firm (originally named Molecular Design Limited, Inc.) in January 1978 in Hayward, California[6].

From its initial pioneering of computer handling of graphical chemical structures with MACCS (Molecular ACCess System) in 1979, MDL continued at the forefront of the field now known as cheminformatics. Innovations include:

- Commercial system for interactive molecular modeling (PRXBLD, 1979)
- Commercial system for computer handling of chemical reactions (REACCS, 1982)
- Commercial system integrating chemical structures with data (MACCS-II, 1984)
- PC-based chemistry database system, the Chemist's Personal Software Series (CPSS: ChemBase, ChemTalk/ChemHost, 1985).

- Commercial system integrating chemistry drawing and word processing (ChemText, 1986)
- Commercial 3D structure database (MACCS-3D, 1988)
- Commercial system for storing structures of mixtures and formulations (MACCS-II Substance Module, 1989)
- Commercial client-server system for managing chemical and biological information (MDL ISIS, 1991)
- Commercial system for structure-based handling of generic structures (MDL Central Library, 1996)
- Searchable, "live" chemical structures in Web pages (MDL Chime plug-in, MDL Chemscape server, 1996)
- Commercial system dynamically linking citations to electronic journal articles and patents (MDL LitLink, 1999)
- Web-based commercial system indexing multiple structure databases and linking to synthetic methodology databases and major reference works (DiscoveryGate, 2002)
- Software application for predicting potential carcinogenic risk of compounds based on their structures (MDL Carcinoge nicity Module, 2003)
- Complete, open, flexible and scalable n-tier informatics system supporting enterprise-wide business process, data and application integration for the life sciences (MDL Isentris, 2004)
- Advanced structure drawing with All-Purpose Drawing Tool and custom template toolbars, MDL Draw, 2004)
- Structure-searchable patent chemistry information (Patent Chemistry Database, 2004) and online pharmacological information (xPharm, 2004)
- Key laboratory workflow applications (MDL Logistics reagent management application and MDL Notebook electronic laboratory notebook, 2005)
- Isentris-based application for building and managing compound registries (MDL Registration, 2006)
- First electronically searchable FDA Drug Approval Packages (PharmaPendium, 2006)
- Out-of-the-box Reaction Planner for exploring precursor steps, integration of proprietary and commercial data (Isentris 2.0 and 3.0, 2007).

Searching

A number of available search modules are provided

- The *standard search* allows querying for systematic names, trade names and synonyms and registry numbers
- The *advanced search* allows interactive searching by chemical structure, chemical substructure, using also molecular formula and molecular weight range, CAS numbers, suppliers, etc. The search can be used to widen or restrict already found results[7].
- The *literature search* allows text-based searching of almost 1/2 million Chemistry Open Access articles from a dozen different sources including the Royal Society of Chemistry, IUPAC, the Journal of Biological Chemistry, the Proceedings of the National Academy of Sciences, the International Union of Crystallography and a number of other publishers[8].

Chemistry document mark-up

The ChemSpider database has been used in combination with text mining as the basis of chemistry document markup. ChemMantis, the Chemistry Markup and Nomenclature Transformation Integrated System uses algorithms to identify and extract chemical names from documents and web pages and converts the chemical names to chemical structures using name-to-structure conversion algorithms and dictionary look-ups in the ChemSpider database. The result is an integrated system between chemistry documents and information look-up via ChemSpider into over 150 data sources[9].

Commercial versus free

The ChemSpider service is owned by the Royal Society of Chemistry, and its services are offered free of charge. Search hits include both free information and pointers into commercial databases that may require a subscription for access[10]. Prior to the acquisition by RSC, ChemSpider was controlled by a private corporation, ChemZoo Inc.

Services

A number of services are made available online. These include the conversion of chemical names to chemical structures, the generation of SMILES and InChI strings as well as the prediction of many physicochemical parameters and integration

to a web service allowing NMR prediction[11]. The organization is working with RSC to develop a hash table resolver for InChIKeys, shorter hashed forms of InChIs.

Refernces:

1. A. Williams, blog post (2007) Chemical & Engineering News. 85(24): 11.

2. Antony Williams (2008) Public Chemical Compound Databases Current Opinion in Drug Discovery & Development. 11(3): 45-78.

3. Bentwich I, Avniel A, Karov Y, et al (2005) Identification of hundreds of conserved and nonconserved human microRNAs. Nat. Genet 37 (7): 766–70.

4. Corey and Wipke (1969) Computer-Asssisted Design of Complex Organic Syntheses. Science 166: 178.

5. Geoff Brumfiel (2008) Chemists Spin a Web of Data. Nature. 453 (7192): 139.

6. Han X, Gross RW (2003) Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: a bridge to lipidomics. J. Lipid Res. 44(6): 1071–9.

7. Kanehisa M. A (1997) Database for post-genome analysis. Trends Genet. 13 (9): 375-6.

8. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res. 32: D277-80.

9. Norton Farman. Chemical & Engineering News (2007) 85(24): 11.

10. Rosenfeld N, Aharonov R, Meiri E, et al. (2008) MicroRNAs accurately identify cancer tissue origin. Nat. Biotechnol. 26 (4): 462–9.

11. Wenk MR. (2005) The emerging field of lipidomics, Nat. Rev. Drug Discov. 4(7): 594–610.

Corresponding author*

Prof. Dr. Dhrubo Jyoti Sen*
Email: *dhrubosen69@yahoo.com*