International Journal of Cheminformatics Research
Vol. 3: Issue 1
www.journalspub.com

Digital Libraries in Chemistry – An Overview

*Shazia Akhtar**
Department of Chemistry, Aligarh Muslim University, Aligarh, Uttar Pradesh, India

ABSTRACT
A chemical digital library is a database specially designed to store chemical data. This information is about chemical and crystal structures, ACD, Zinc, MayBridge, MedChem, Beilstein, WDI, WOMBAT, TSCA, and Thermophysical data. This review paper presented the overview of different types of digital libraries/databases uses in chemistry.

Keywords: ACD, Beilstein, MayBridge, MedChem, thermophysical data

***Corresponding Author**
E-mail: a.shazia771900@gmail.com

INTRODUCTION

Cheminformatics is a combination of chemistry and information technology, is required for the processing and analysis of chemical data. Cheminformatics is relevant to biologists because chemistry data are important in many areas of molecular biology, e.g., in the study of protein interactions and metabolism. Structural information about different molecules can be obtained from a number of comprehensive resources/digital libraries/databases, including ACD, Zinc, MayBridge, MedChem, Beilstein, WDI, WOMBAT, TSCA, and Thermophysical data. Each of these resources provides a chemical database that can be searched using a variety of query formats, e.g., systematic name, non-systematic name, formula, molecular weight or CAS registry number. Search results provide physical, chemical and biomedical information with links to other databases and resources. Med hem also provides the SMILES string [1].

Chemical Digital Library

Chemical Digital libraries are required for efficient lead discovery if little is known about the binding properties of the drug target. Conversely, focused libraries are required if the structure of the target is known, since this defines a particular set of ligands. Chemical diversity can be defined by comparing molecules on the basis of descriptors (functional groups) and how these fill chemical space. A number of software tools are available for the design and assessment of diverse or focused chemical libraries, virtual screening against drug targets [1].

TYPES OF CHEMICAL DIGITAL LIBRARIES/DATABASES

Chemical Structures

Chemical structures are traditionally represented using lines indicating chemical bonds between atoms and drawn on paper (2D structural formulae). While these are ideal visual representations for the chemist, they are unsuitable for computational use and especially for search and storage. Small molecules (also called ligands in drug design applications), are usually represented using lists of atoms and their connections. Large molecules such as proteins are however more compactly represented using the sequences of their amino acid building blocks. Large chemical databases for structures are expected to handle the storage and searching of information on

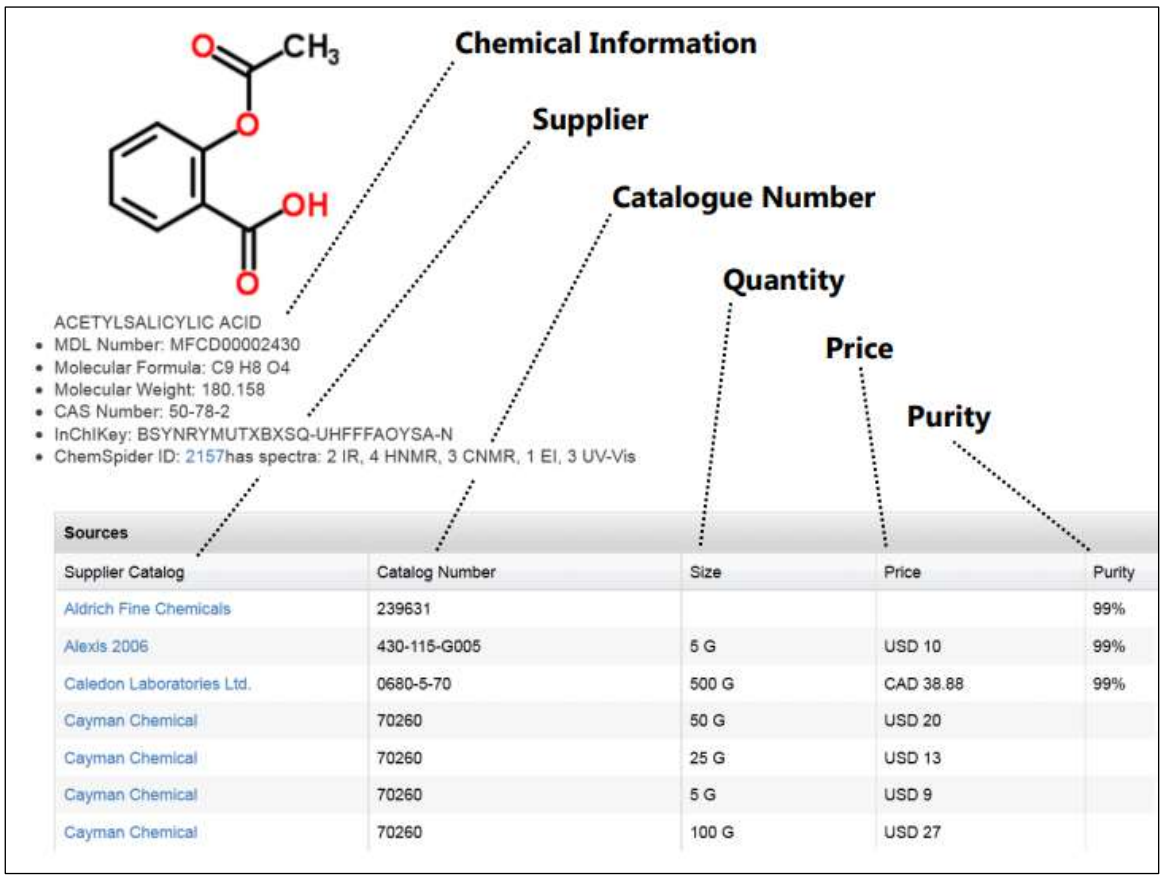
millions of molecules taking terabytes of physical memory [2].

Literature Database

Chemical literature databases correlate structures or other chemical information to relevant references such as academic papers or patents. This type of database includes STN, Scaffolder, and Reaxys. Links to literature are also included in many databases that focus on chemical characterization [2].

Available Chemical Directory (ACD)

The available chemicals directory (ACD) is a database of commercially available chemicals that can be searched by structure. Pricing and supplier information is provided for 3.2 million unique chemicals from over 800 suppliers (Figure 1). The ACD can be searched by substance name or structure / substructure. The Available Chemicals Directory is provided to the UK academic community via the Royal Society of Chemistry-hosted Chemical Database Service at cds.rsc.org. The Available Chemicals Directory has been developed by Accelrys. The Chemical Database Service is funded by the EPSRC [3].



Supplier Catalog	Catalog Number	Size	Price	Purity
Aldrich Fine Chemicals	239631			99%
Alexis 2006	430-115-G005	5 G	USD 10	99%
Caledon Laboratories Ltd.	0680-5-70	500 G	CAD 38.88	99%
Cayman Chemical	70260	50 G	USD 20	
Cayman Chemical	70260	25 G	USD 13	
Cayman Chemical	70260	5 G	USD 9	
Cayman Chemical	70260	100 G	USD 27	

Fig. 1. *Information provided by ACD.*

ZINC (UCSF)

ZINC, a free database of commercially-available compounds for virtual screening. ZINC contains over 35 million purchasable compounds in ready-to-dock, 3D formats. ZINC is provided by the Irwin and Shoichet Laboratories in the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF). ZINC includes over 400 catalogs from over 300 vendors and over 100 annotated catalogs. ZINC is widely used. They have 500 unique

visitors per day, 13,000 per month, and have over 50,000 "repeat customers" [4].

ZINC was originally designed for target based virtual screening (docking), and this remains its primary focus. However, ZINC is also useful for many other things, including:

- Finding a compound to purchase
- Downloading a library in SMILES format for ligand based virtual screening
- Find compounds by similarity to a starting compound (SAR-by-catalog)
- Find compound ANNOTATED for a particular target (via ChEMBL)
- Find compounds PREDICTED for a particular target (via SEA/ChEMBL or docking)

ZINC is updated continuously. In a typical week:

- 100,000 new molecules are loaded
- 10,000 molecules are repaired
- 80,000 catalog items are marked "depleted" due to their absence from the most current catalogs.
- 3 new catalogs are added.
- 30 tranches of the 2D and 3D property subsets are updated.

MayBridge

The Maybridge portfolio offers a comprehensive range of chemistry products and services tailored to the drug discovery and biotechnology sector. For over 50 years, Maybridge has been at the forefront of innovative heterocyclic building block and screening compound design, fueled by the desire to access novel molecules of pharmaceutical interest. The Maybridge portfolio is driven by a keen understanding of the needs of the medicinal chemist and is designed to expedite the drug discovery process. Maybridge specializes in producing innovative heterocyclic chemical building blocks for drug discovery chemistry. Building Blocks collection is a unique and expanding range of Reactive Intermediates, sets of minimally substituted building blocks sharing a common ring structure, each functionalized with a selection of the most synthetically useful reactive groups. Minimal substitution means easier interpretation of SAR in lead optimization, and the diversity of functional groups attached to each ring structure allows the chemist maximum flexibility in library design and production. Reactive Intermediates represent just a subset of our vast range of structurally and chemically diverse building blocks, which provides the ultimate toolbox for drug discovery chemistry. The drug discovery process is long and expensive. Maybridge aim is to shorten this process by producing high-quality, hit-like, lead-like and drug-like compounds, which generate quality valuable data from screening programmed. The Maybridge Screening collection consists of over 53,000 organic compounds, largely produced by us at Maybridge. These are individually designed compounds, produced by innovative synthetic techniques, based on over 45 years of experience in heterocyclic chemistry [5].

MedChem

This database is published by BioByte Corp. and Pomona College under the category Pharmaceutical database. It is updated annually. The version available in this database is 2003.

The Medchem Project and Biobyte Corp. in Claremont. CS, maintains this database. Medchem03 consist of 48,500 compounds, 67,000 measured logP's, 13700 pKa's, 65,000 names. 30,500 CAS numbers, 19,000 activities, CLOGP values for 46,000 compounds, and Rubicon 3D coordinates for 43,000 structures.

The Medchem database is available in Daylight's TDT format for THOR/Merlin users. The Medchem database is licensed on an annual basis. Periodic updates are included with the subscription.

The MedChem and BioByte Corp. in Claremont, CA, maintain the MedChem database. This database contains more than 55,000 compounds with 75,000 names and 32,000 CAS numbers. In addition, there are 61,000 measured logP values, 13,900 pKa values, 26,000 activities, and Rubicon 3D coordinates for almost 49,000 structures [6].

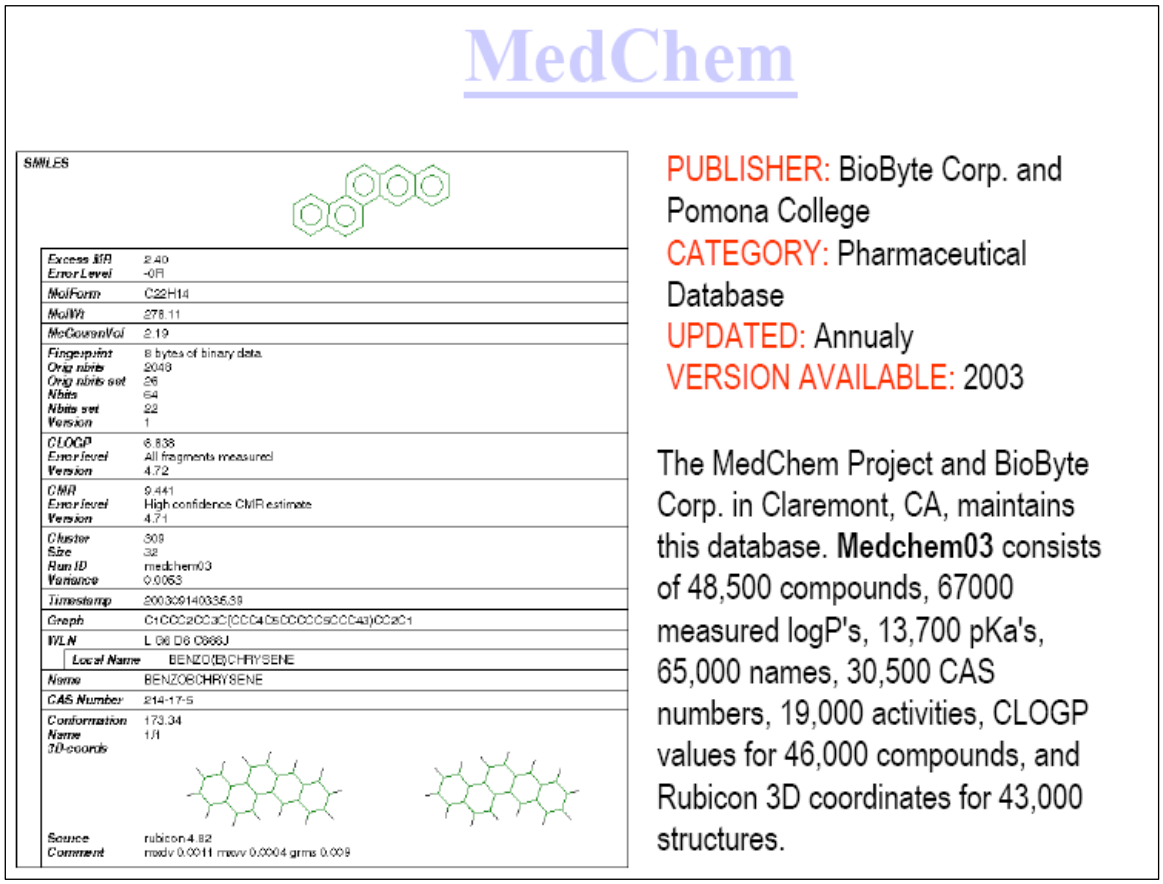


Fig. 2. *MedChem database.*

Beilstein

The main focus of Beilstein is on compound and reactions, rather than citations. Compounds can be broken down by chemical type, with approximately 53% heterocyclic, 37% isocyclic and 9% acyclic, the remaining 1% including mixtures, polymers and biomolecules. A major component of the database is reaction information. Beilstein contain approximately 107 reactions over half of which have graphical representations and are fully searchable by construction of structural queries through a structural drawing system. Beilstein is the largest organic reactions database in existence. With more than 35 million entries in several hundred fields, Beilstein covrs chemical, physical and biological properties. For paper published since 1980, author abstracts are also included; approximately 750,000 are now available in this database. The chemical literature dating from 1771 is also included [7].

World Drug Index (WDI)

Derwent Publications maintains this drug database of almost 80,000 drugs and pharmacologically active compounds, including all marketed drugs. WDI also

contains over 200,000 synonyms, 78,000 trade names, 7,300 International Nonproprietary Names, 8,300 US Adopted Names, 28,000 journal and conference references, and more, including manufactures and extensive medical data, such as indications and usage, interactions, adverse effects, mechanism of action, and activity keywords. WDI is available in DayCart®/Oracle import format and in Daylight's TDT format for THOR/Merlin users. WDI is licensed on per-user and on an annual subscription basis [8].

World of Molecular Bioactivity (WOMBAT)

The database is published by Sunset Molecular Discovery under the Category–Molecule database. It is updated twice yearly. Version Available for this database is 2004. WOMBAT is a database of biologically active compounds. Sunset molecular discovery maintains the database of over 70,000 entries, including over 68,000 unique SMILIES, over 3000 papers, covering over 140,000 activities on 630 targets (Figure 3). WOMBAT contains

information published in the following journals from 1975–2003: *Biochem. Pharmacol, Bioorg. Med. Chem. Lett., Chembiochem, Eur. J. Med. Chem., J. Amer. Chem Soc., J. Med. Chem, Quant Struct–Act. Relat and J. Healt Sci [9].*

Toxic Substance Control Act (TSCA)

The database is published by US Environmental Protection Agency. Version Available in this database is 1993 (Figure 4). This database is comprised of substances in the "Discrete Organic Subsel" of the US toxic substance Control Act. Over 100,000 substances are represented, of which approximately 39,000 have associated SMILES. There are approximately 78,000 CAS numbers, 61,000 "Preferred names", 46,000 "Submitter names", and more [10].

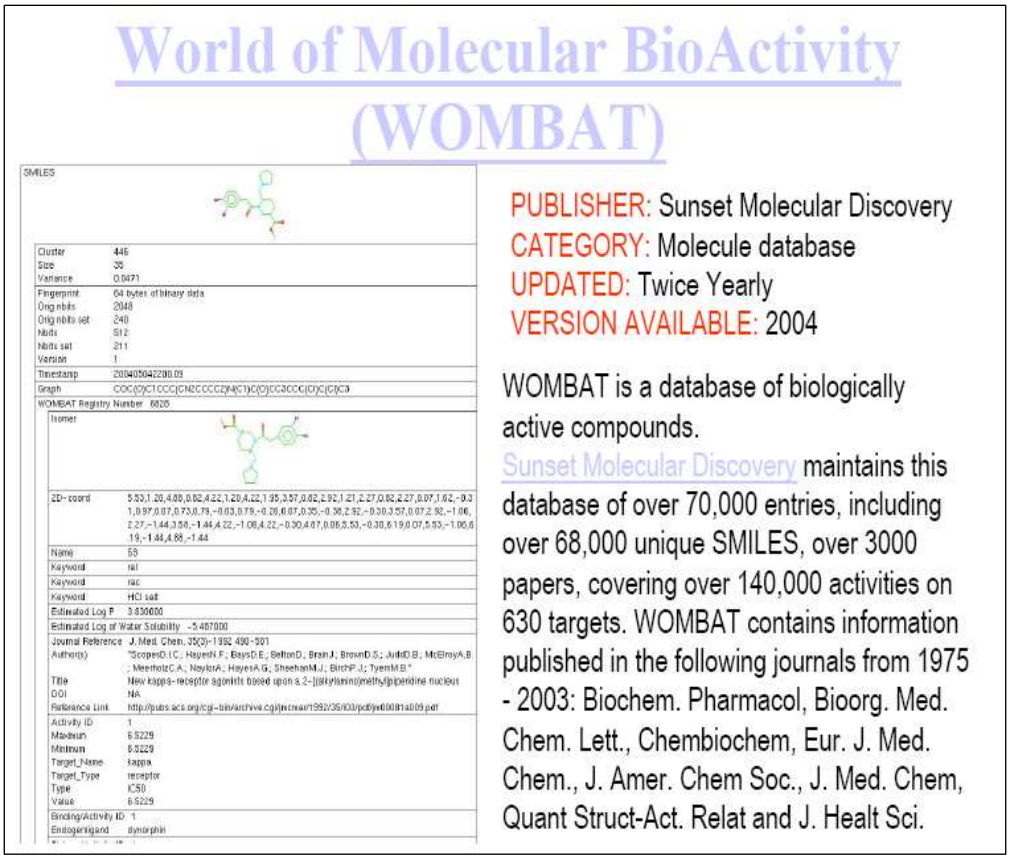


Fig. 3. *WOMBAT database.*

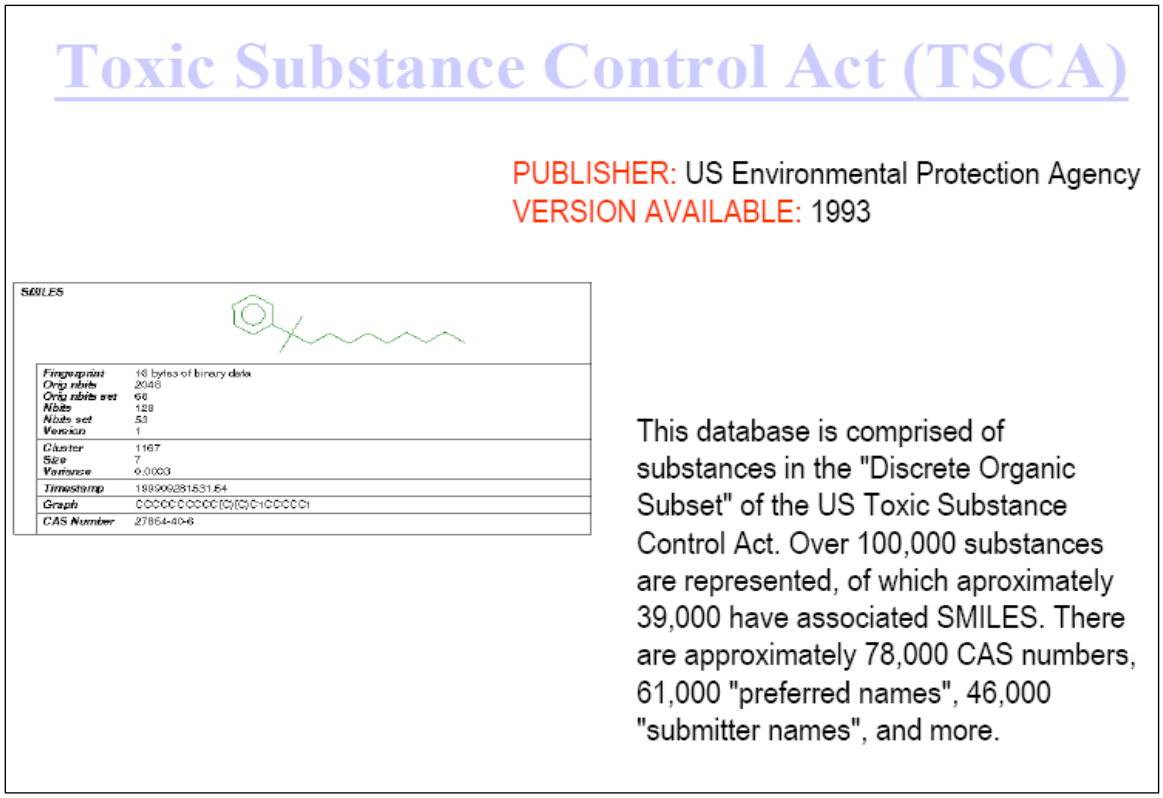


Fig. 4. *TSCA database.*

Thermophysical Database

Thermophysical data are information about phase equilibria including vapor–liquid equilibrium, solubility of gases in liquids, liquids in solids (SLE), heats of mixing, vaporization, and fusion. Caloric data like heat capacity, heat of formation and combustion, transport properties like viscosity and thermal conductivity [11].

Chemical Structure Representation

There are two principal techniques for representing chemical structures in digital databases.

As connection tables/adjacency matrices / lists with additional information on bond (edges) and atom attributes (nodes), such as:

- MDL Molfile, PDB, CML
- As a linear string notation based on depth first or breadth first traversal, such as:
- SMILES/SMARTS, SLN, WLN, InCh I
- These approaches have been refined to allow representation of stereochemical differences and charges as well as special kinds of bonding such as those seen in organo-metallic compounds. The principal advantage of a computer representation is the possibility for increased storage and fast, flexible search [12].

Tools

The computational representations are generally made transparent to chemists by graphical presentation of the data. Data entry is also cut down through the use of chemical structure editors. These editors internally transform the graphical data into computational representations. There are also several algorithms for the interconversion of numerous formats of representation. An open-source utility for conversion is OpenBabel. These search and conversion algorithms are applied either within the database system itself or as is now the trend is executed as external components that fit into standard relational

database systems [13]. Both Oracle and PostgreSQL based systems make use of cartridge technology that allows user defined datatypes [14]. These permit the user to make SQL queries with chemical search conditions (For example, a query to search for records having a phenyl ring in their structure represented as a SMILES string in a SMILESCOL column could be

```
SELECT * FROM CHEMTABLE
WHERE
SMILESCOL.CONTAINS('c1ccccc1')
```

Algorithms for the conversion of IUPAC names to structure representations and *vice versa* are also used for extracting structural information from text. However, there are difficulties due to the existence of multiple dialects of IUPAC. Work is on to establish a unique IUPAC standard.

CONCLUSION

Chemical Digital Libraries could play a greater role for researchers, scientists and professor. Libraries/databases are designed in such a way that researchers improve the quality of their funding applications, and to increase the institution's success in winning research income. Chemical digital Libraries are critically important in helping researchers to solving out their research related queries, information etc. The digital revolution has changed the relationship between libraries and researchers. Most research institutions now have chemical digital libraries to store and make available chemical assets for examples chemical structures, chemical reactions, research papers and theses, etc. Chemical Digital Library is now playing an increasing role in educating researchers and building more effective procedures and approaches across the institution. The value of the digital library is as a crucial cornerstone for all the researchers.

REFERENCES

[1] J.R. Ullmann. An algorithm for subgraph isomorphism, *J ACM*. 1976; 23(1): 31–42p.
[2] S.A. Rahman, M. Bashton, G.L. Holliday, R. Schrader, J.M. Thornton. Small molecule subgraph detector (SMSD) toolkit, *J Cheminform*. 2000; 1: 12p.
[3] http://blogs.rsc.org/chemical-database-service/2013/10/31/available-chemicals-directory/?doing_wp_cron=1502350503.9696118831634521484375.
[4] http://zinc.docking.org/.
[5] http://www.maybridge.com/.
[6] M.D. Cummings, A.C. Maxwell, R.L. DesJarlais. Processing of small molecule databases for automated docking, *Med Chem*. 2007; 3(1): 107–13p.
[7] Reaxys. Retrieved 2013-03-17.
[8] "Reaxys". Retrieved 4 January 2011.
[9] http://www.daylight.com/meetings/mug97/GCross/wdi_htdm.htm.
[10] https://www.researchgate.net/publication/229618698_Chemical_Informatics_WOMBAT_and_WOMBAT-PK_Bioactivity_Databases_for_Lead_and_Drug_Discovery.
[11] https://www.epa.gov/laws-regulations/summary-toxic-substances-control-act.
[12] http://www.infotherm.com/.
[13] S.A. Rahman, M. Bashton, G.L. Holliday, R. Schrader, J.M. Thornton. Small molecule subgraph detector (SMSD) Toolkit, *J Cheminform*. 2009; 1: 12p.
[14] D. Butina. Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: a fast and automated way to cluster small and large data sets, *Chem Inf Comput Sci*. 1999; 39: 747–50p.