International Journal of Current Pharmaceutical Research

ISSN- 0975-7066 Vol 3, Issue 1, 2011

Review Article

CHEMOINFORMATICS: A NOVEL TOOL IN DRUG DISCOVERY

V.S.VELINGKAR, GAURAV POKHARNA*, NITIN S. KOLHE

*Prin. K. M. Kundnani College of Pharmacy, Department of Pharmaceutical Chemistry, 23, Jote Joy Bldg, R.S.Marg, Cuffe Parade, Colaba, Mumbai 400005, India.Email: gaurav_pokharna@rediffmail.com

Received: 14 Aug 2010, Revised and Accepted: 18 Sep 2010

ABSTRACT

The discovery of new chemical entities exhibits a paradigm shift by application of novel techniques like combinatorial chemistry and high-throughput screening generating huge amount of data. This data and information can only be managed and made accessible by storing them in databases. Such problems in chemistry require use of chemoinformatics methods. Chemoinformatics is the application of informatics methods to solve chemical problems. It covers the application of computer-assisted methods to chemical problems like information storage and retrieval, the prediction of physical, chemical or biological properties of compounds, spectra simulation, structure elucidation, reaction modeling, synthesis planning and drug design. Chemoinformatics methods have successfully been applied in all fields of chemistry. The future will bring a rapid expansion of the use of Chemoinformatics to our further understanding of chemistry and to process the flood of chemical information.

Keywords: Chemoinformatics, Graph theory, SMILES, Similarity, Bioinformatics.

INTRODUCTION

Chemistry is largely built on experimental observations and data, it deals with compounds, their properties and their transformations. Compounds and chemical reactions are the static and dynamic aspects of chemistry. The entire living and material world consists of compounds and mixtures of compounds. Compounds are transformed into each other by chemical reactions that can be run under a variety of condition. Although the laws of chemistry are too complicated to be solved, chemists still can do their jobs and make compounds with beautiful properties that society needs, and chemists still run reactions from small-scale laboratory experiments to large scale reactors in chemical industry. The secret to success has been to learn from data and from experiments. The process of learning is called inductive learning as shown in Fig. 1.

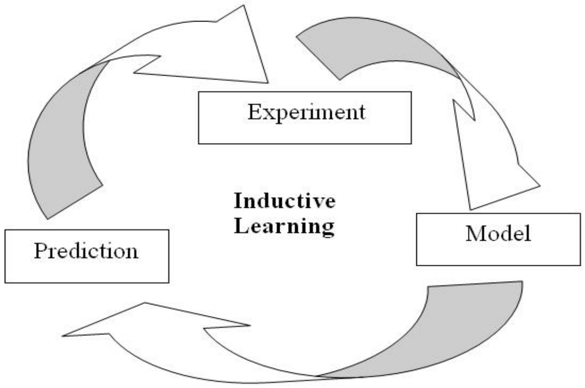

Fig. 1: Inductive Learning Model

Amount of data and information is enormous and increasing rapidly. At present, more than 41 million different compounds are known, all have a series of properties, physical, chemical and biological; all can be made in many different ways, by a wide range of reactions; all can be characterized by a host of spectra. Problem is to extract knowledge from these data and use it to make predictions.

Three major tasks of structure-property / activity relationships, design of reaction / syntheses and structure elucidation are tackled by making use of prior information, and of information that has been condensed into knowledge. The amount of information that has to be processed is often quite large. This immense amount of information can be processed only by electronic means, by the power of computer. This is how chemoinformatics is useful[1].

Definitions

"The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization."- F.KBrown[2]

"Chemoinformatics - a new name for an old problem. "- M. Hann, R Green[3]

"Chem(o)informatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization and use of chemical information."- G. Paris[4]

Chemoinformatics is the application of informatics methods to solve chemical problems. - J.Gastegeir et.al[5]

History

There is no particular point in time that determines when chemoinformatics was founded or established. It slowly evolved from several, often quite humble beginnings. Scientists in various fields of chemistry struggled with the development of computer methods, which allowed them to manage the enormous amount of chemical information and to find relationships between the structure and properties of a compound. During the 1960s some early developments appeared that led to a flurry of activities in the 1970s. The sequence of events occurred in order to arrive at what is called as chemoinformatics today are as follows: The first information systems and services were paper-based (*Annalen der Pharmacie* founded in 1832 and *Chemical Abstracts* started in 1907).The first Computer-based systems established over 40 years ago. The first journal for chemical information, Journal of chemical documentation was started in 1961 (name changed to Journal of Chemical Information and Computer science in 1975).The first book, Computer Handling of Chemical Structure Information was appeared in 1971.The first international conference on topic was held in 1973 at Noordwijkerhout[8-11].The term Chemoinformatics was defined by F.K. Brown in 1998[2].

Objectives

Chemoinformatics should assist the chemist to solve some of following fundamental problems:

1. To design molecules with desired properties - The major task of a Chemist is to make compounds with desired properties, establish structure-activity or structure-property relationships (SAR or SPR) or even of finding such relationships in a quantitative manner (QSAR or QSPR).

2. To design reaction and syntheses to make these compounds - The designing of reaction includes the sequence of reactions and starting materials to be used to synthesize the desired compound.

3. To analyze and elucidate the structures obtained in reactions - There is a need to establish the structure of the reaction product by using various tools of structure elucidation.

4. To transform data into knowledge through information processing for the intended purpose of making better decisions faster[1]. Fig.2. depicts the pyramid from data through information to knowledge[2].

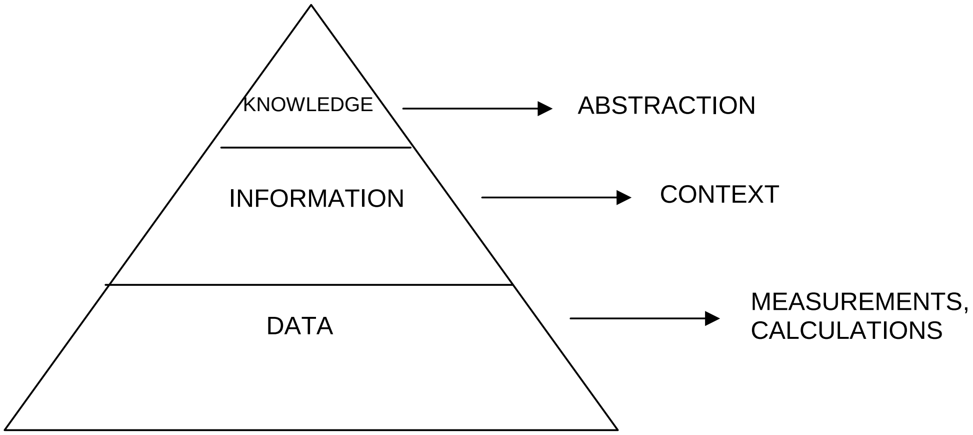

Fig.2. From data through information to knowledge[2]

Need

1. Chemical information explosion: Chemical Abstracts Service adds over three-quarters of a million new compounds to its database annually, for which large amounts of physical and chemical property data are available. Some groups generate hundreds of thousands to millions of compounds on a regular basis through combinatorial chemistry that are screened for biological activity. Even more compounds are generated and screened *in silico* in the search for a magic bullet for a given disease. Combinatorial chemistry and high-throughput screening are data dependent and data rich technologies. When making combinatorial libraries of chemical compounds, you need information on the molecular components, their biological effects, and information on how to prepare the compound. There is also data for managing and storing the libraries. In high throughput screening, the test results need to be captured, stored, and then analyzed[6-7].

2. Three dimensional structures determined by x -ray crystallography known for about 300,000 organic compounds. Or the largest database of infrared spectra contains about 200,000 spectra. It is only 1% of all the available compounds. The question is then; can we gain enough knowledge from the known data to make predictions for those cases where the required information is not available? This is another reason why we need informatics methods in chemistry[5].

3. Many problems in chemistry are too complex to be solved by methods based on inductive learning or through theoretical calculations. This is true, for the relationships between the structure of a compound and its biological activity, or for the influence of reaction conditions on chemical reactivity. All these problems in chemistry require novel approaches for managing large amounts of chemical structures and data, for knowledge extraction from data, and for modeling complex relationships.

This has created a demand to collect, organize, and apply the chemical information. This is where chemoinformatics methods come in[1].

Representation OF 2d molecular structures

Structures are needed to be included in computer readable form. Emphasis laid on computational representation of molecular structures and creation of structural databases.

Different ways by which structures can be represented are: 1) Image file, 2) Graph theory, 3) Connection tables and 4) Linear notation.

Image file: easy-to-use computer programs such as *Chemdraw* & *ISIS/Draw* represent Structures. It enables chemists to draw structures for incorporation into reports, publications or presentations. One way to store the structure would be as an image, as might for example be obtained by scanning a printout of the structure into the computer, or in a text format such as PDF. For retrieval and search through database it is not favorable. Fig.3. shows the image file of Aspirin.

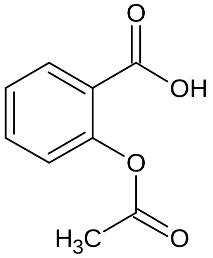

Fig. 3: Aspirin

Graph theory: chemical structures are usually stored in a computer as molecular graphs. A graph is an abstract structure that contains nodes connected by edges. In a molecular graph the nodes correspond to the atoms and the edges to the bonds. Hydrogen atoms are often omitted. A graph represents the topology of a molecule only, that is, the way the nodes (of atoms) are connected. Acyclic molecules are represented using trees. For example Aspirin graph in Fig.4.

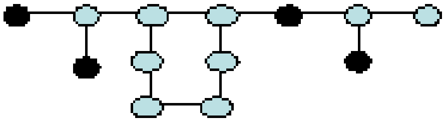

Fig. 4: Aspirin graph

Connection tables: it is a means to communicate the molecular graph to and from the computer. The simplest type of connection tables consists of two sections: first, a list of the bonds, specified as pairs of bonded atoms. More detailed form of connection table includes: List of atomic numbers, List of the bonds, Hybridization state, Bond order, Information about xy or xyz coordinates of atoms.

Linear/line notations: it represents the structure of chemical compounds as a linear sequence of letters and numbers. They are more compact than connection tables, so useful for storing & transmitting large no. of molecules. Various line notations are: WLN[12] (Wiswesser Line Notation), ROSDAL (Representation of Organic Structures Description Arranged Linearly), SMILES[13] (Simplified Molecular Input Line Entry Specification) and SLN (Sybyl).Out of above-mentioned notations SMILES are widely accepted. In SMILES, atoms are represented by their atomic symbols, Upper case – for aliphatic, Lower case – for aromatic, Hydrogen - not represented, Double bonds – "=" , Triple bonds - "#", Single and aromatic bonds are not explicitly represented by any symbol. Examples: Cyclohexane – C1CCCCC1, Benzene - c1ccccc1, 2-Methyl propane – CC(C)C, Acetic acid – CC(=O)O[7].

Canonical representation of molecular structures

Canonical representation is unique ordering of atoms for a given graph. It is required due to different ways of constructing connection table or the SMILES string for a given molecule. In a connection table one may choose different ways to number the atoms and in SMILES notation the SMILES string may be written starting at a different atom or by following a different sequence through the molecule. A well-known and widely used method for determining a canonical order of the atoms is the Morgan algorithm[14] and SEMA. An algorithm called CANGEN[15] has been developed to generate a unique SMILES string for each molecule. For example canonical representation of Aspirin is: CC (=O) Oc1ccccc1C (=O) O.

Representation of three dimensional (3D) molecular structures

Two Dimensional (2D) representations of molecules only tell about atoms, which are bonded together. It doesn't tell about steric & electronic parameters and atom positions in 3D space. Three Dimensional representations of molecules have following challenges: 1) Molecules can adopt more than one low energy conformation and 2) The number of accessible structures is very large. So there is need to represent molecular structures in 3D. The data stored in a 3D database either comes from Experimental methods or Computational methods[7].

Experimental 3D databases: It includes structures solved using X-ray Crystallography. The CSD[16] (Cambridge Structural Database) contains the X-ray structures of more than 250,000 organic and organometallic compounds. The Cambridge Structural Database (CSD) stores crystal structures of small molecules and provides a fertile resource for geometrical data on molecular fragments for calibration of force fields and validation of results from computational chemistry[17-18].

As protein crystallography gained momentum, the need for a common repository of macromolecular structural data led to the Protein Data Base (PDB) originally located at Brookhaven National Laboratories[19]. The PDB (Protein Data Bank) contains more than 20,000 X-ray & NMR structure of proteins and protein-ligand complexes and some nucleic acid and carbohydrate structures. Both these databases are widely used and continue to grow rapidly.

Theoretical 3D databases: The PDB & CSD are extremely useful but for most compounds no crystal structure is available. There is also an increasing need to evaluate virtual compounds – structures that could be synthesized but which have not yet been made. Even when experimental data is available, this usually provides just a single conformation of the molecule which will not necessarily correspond to the active conformation; most molecules have a number of conformations accessible to them.

It is thus desirable to include mechanisms for taking conformational space of the molecules into account during 3D database searching. Structure-generation programs such as CONCORD[20], CORINA[21], and COBRA take 2D representation of molecule and generate a low energy conformation. These programs generate one conformation. More conformation can be generated by tweaking or by storing multiple conformations.

Data and databases

Data refers to a collection of organized information, usually the results of experience, observation or experiment, or a set of premises. This may consist of numbers, words, or images, particularly as measurements or observations of a set of variables. Data can be categorised into four types:

1) Structural data – it refers to the 1-, 2- or 3-D representations of molecules.

2) Numerical data – it includes biological activity, pka, log P, or analytical results

3) Annotation/text – it includes information such as experimental notes that are associated with a structure or data point.

4) Graphical data – any structure or data point may have associated graphical information such as spectra or plots.

In all cases the data may be experimental or computed and the molecules may be real or virtual. The Internet is increasingly used to distribute data and information in chemistry[2, 22].

Database is defined as a self-describing collection of integrated records, mainly stored on hard disk or CD – ROM.

Classification of databases

Databases can be classified into following three type[22]:

Literature databases – author names, titles, journals or books etc. can be stored & retrieved. It is of three types: Bibliographic (example CA file, Medline, SCISEARCH, BIOSIS), Full text (electronic journals example ACS), and Patent (example MARPAT[23], INPADOC, WPINDEX).

Factual databases – contains alphanumeric data or compounds. It is of four types: Numeric (example. DETHERM, Web Book, Spec Info, and CSD), Metadata (example Gule directory of databases), Directory (example FEDRIP, NUMERI-GUIDE, UFORDAT), and Catalogs (example Chemline & MRCK)

Structure databases – contains information on chemical structure and compounds. It is of two types: Structure (example Beilstein[24], Gmelin, CAS registry), and Reaction (example ChemInformRX, CAS REACT[25], ChemReact).

Searching chemical structures

Full structure search: It includes the searching as well as retrieval of information from databases. Structure searching can be done by use of molecular formula, molecular weight, Trade and/or trivial name, registry number (CAS & Beilstein), and hash codes.In it, structure is first converted to canonical representation and then Hash code[26] is generated. Hash code corresponds to physical location on the computer disk. A hash code is a fixed length representation of a data structure used as an index or key to a direct access file. The input structure cannot be restored from a hash code, and due to the limited range of values two different data structures may be represented by the same hash code. Ihlenfeldt and Gasteiger proposed to represent chemical structures with hash codes by using a hierarchical algorithm: atom hash codes are computed first, merged into molecule hash codes and the molecule hash codes are combined to give a molecular ensemble hash code[27].

Substructure search: Substructure search identifies all the molecules in the database that contain a specified substructure. Two methods are used for substructure search, they are 1) Sub graph isomorphism[28]: to determine whether one graph is entirely contained within another. It is a slow process and has factorial degree of complexity. 2) Bitstrings: it consist of a sequence of "0"s and "1"s. A "1" in a bitstring usually indicates the presence of a particular structural feature and a "0" its absence. Fig.5. 2D Substructure searching.

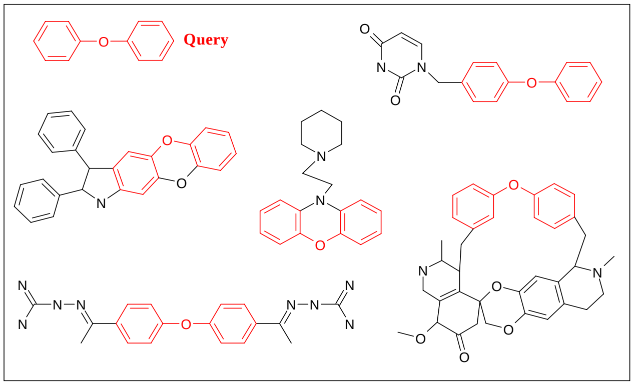

Fig. 5: 2D Substructure searching

Three dimensional structure search: Its objective is to identify conformations that match the query. It is a two-stage process[29]: a) Rapid screen to encode information about the distances between relevant groups in molecular conformation b) a subgraph isomorphism such as Ullmann algorithm[30]. Potential matches are then confirmed by fitting the relevant conformation to the query in Cartesian space. Fig.6. 3D Substructure searching.

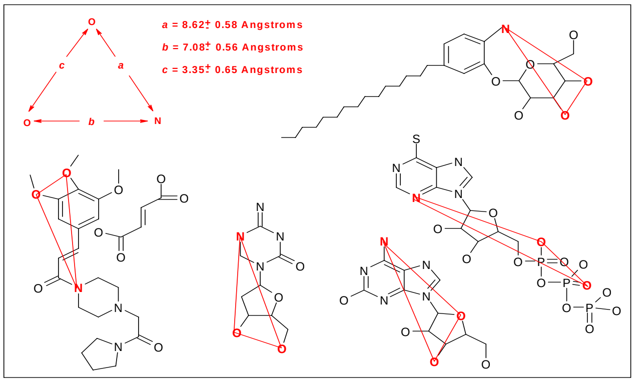

Fig.6. 3D Substructure searching

Molecular similarity and molecular diversity

The concepts of molecular similarity[31-33] and molecular diversity[34-35] play important roles in modern approaches to Computer Assisted Molecular Design. Molecular similarity provides the simplest and most widely used method for virtual screening and underlies the use of clustering methods on chemical databases. Molecular diversity analysis provides a range of tools for exploring the extent to which a set of molecules spans structural spaces and underlies many approaches to compound selection and to the design of combinatorial libraries.

Similarity- Property Principle: The principle states that structurally similar molecules are expected to exhibit similar properties. It is clear that there are many exceptions to the principle as stated[36-37], because even a small change in the structure of a molecule can bring about a radical change in some property.

The concept of similarity is important and many different similarity coefficients have been developed. Examples of different Similarity Coefficients (S_{AB}) are:

1. Tanimoto Coefficient $S_{AB} = \dfrac{c}{a+b-c}$

2. Cosine Coefficient $S_{AB} = \dfrac{c}{\sqrt{ab}}$

3. Hamming Distance $S_{AB} = [a + b - 2c]$

Where, "a" & "b" bits set to 1 in molecule A & B respectively "c" bits set to 1 common

to both A &B. The standard approach is to use 2D fragment bit-strings in conjunction with the Tanimoto Coefficient[38].

Applications

The range of applications of chemoinformatics is rich indeed; any field of chemistry can profit from its methods. The following lists different areas of chemistry and indicates some typical applications of chemoinformatics. It has to be emphasized that this list of applications is by far not complete.

Chemical Information

- Storage and retrieval of chemical structures and associated data to manage the flood of data.
- Dissemination of data on the internet.
- Cross-linking of data to information.

All fields of chemistry

- prediction of the physical, chemical, or biological properties of compounds

Analytical Chemistry

- Analysis of data from analytical chemistry to make predictions on the quality, origin, and age of the investigated objects.
- Elucidation of the structure of a compound based on spectroscopic data

Organic Chemistry

- Prediction of the course and products of organic reactions.
- Design of organic syntheses.

Drug Design

- Identification of new lead structures.
- Optimization of lead structures.
- Establishment of quantitative structure-activity relationships.
- Comparison of chemical libraries.
- Definition and analysis of structural diversity.
- Planning of chemical libraries.
- Analysis of high-throughput data.
- Docking of a ligand into a receptor.
- de novo design of ligands.
- Modeling of ADME-Tox properties.
- Prediction of the metabolism of xenobiotics.
- Analysis of biochemical pathways.

Bioinformatics

- It generally focuses on genes & proteins, while chemoinformatics centers on small molecules.
- For proteins to perform function there is a need to maintain the specific 3D structure. This evolutionary history is used successfully for aligning proteins (or nucleotide) sequences. Generally advanced alignment algorithms use programs such as BLAST[39] and FASTA[40] and then apply dynamic programming algorithm[5].

There are many areas and problems that can still benefit from the application of chemoinformatics methods. There is much space for innovation in seeking for new applications and for developing new methods[1, 5, 7].

CONCLUSION

Chemoinformatics has developed over the last 40 years to a mature discipline that has applications in any area of chemistry. As high-throughput technologies and combinatorial chemistry continue to advance, informatics techniques will become indispensable in managing and analyzing the exploding volumes of data. By organizing, the data with the help of Chemoinformatics will catalyze further advancements and open new possibilities in the field of drug discovery. There are still many problems that await a solution and therefore many new developments in chemoinformatics are foreseen.

Future prospects

Chemoinformatics will gain importance in chemistry and is incorporated into regular chemistry curricula. Use of computer assisted Structure Elucidation (CASE) process and Computer Assisted Synthesis Design (CASD) would be integrated into the daily work process of bench chemists. Chemoinformatics methods will be extended to theoretical chemistry, simulation of reactions, modeling of biochemical and metabolic reaction, study of proteins will be the future areas of thrust for chemoinformatics. Another field of great activity will be the merging of bioinformatics and chemoinformatics; their common problems can be solved using methods developed in both the fields[5].

REFERENCES

1. Gasteiger J, Engel T, editors.Chemoinformatics - A Textbook. Germany: Wiley-VCH; 2003.
2. Brown FK. Chemoinformatics: what is it and how does it impact drug discovery. Annual Reports of Medicinal Chemistry. 1998; 33: 375-384.
3. Hann M, Green R.Chemoinformatics– a new name for an old problem? Curr. Opin. Chem. Biol.1999; 3:379–383.
4. Paris G. (August 1999 Meeting of the American Chemical Society), quoted by W.Warr at http://www.warr.com/warrzone.htm.
5. Gasteiger J, editor. Handbook of Chemoinformatics-From Data to Knowledge. Germany: Wiley-VCH; 2003.
6. Russo E. Chemistry Plans a Structural Overhaul. Nature. 2002; 419: 4 –7.
7. Leach AR, Gillet VJ. An Introduction to Chemoinformatics. Netherlands: Kluwer Academic Publishers; 2003.
8. Nourse JG et.al. J Chem Inf Comput Sci.1988; 32- 34.
9. Weininger D. J Chem Inf Comput Sci.1988; 28: 31-37.
10. Weininger D. J Chem Inf Comput Sci. 1989; 29: 97-98.
11. Vishwanandhan VN, Ghose AK, Revankar GR et.al. J Chem Inf Comput Sci.1989; 29: 163-169.
12. Wiswesser WJ. A Line – Formula Chemical Notation. New York: Corwell Co; 1954.
13. Weininger D. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. J Chem Inf Comput Sci. 1989; 29: 97 – 101.
14. Morgan HL. The Generation of a Unique Machine Description for Chemical Abstracts Service. J Chem Documentation. 1965; 3: 107 – 113.
15. Weininger D, Weininger A, Weininger JL. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. J Chem Inf Comput Sci. 1989; 29: 97 – 101.
16. Allen FH, Bellard SA, Brice MD et.al. The Cambridge Crystallographic Data Centre: Computer Based Search, Retrieval, Analysis and Display of Information. Acta Crystallographica. 1979; B35: 2331-2339.
17. Allen FH. The Cambridge Structural Database: A Quarter of a Million Crystal Structures and Rising. Acta Crystallographica. 2002; B58: 380-388.
18. Bernstein FC, Koetzle TF, Williams GJB et.al. The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. J Molecular Biology. 1977; 112: 535-542.
19. Berman HM, Westbook J, Feng Z et.al. The Protein Data Bank. Nucleic Acids Research. 2000; 28: 235-242.
20. Rusinko A III, Skell JM, Balducci et.al. CONCORD: A Program for the Rapid Generation of High Quality 3D Molecular Structures. The University of Texas at Austin and Tripos Associates: St. Louis Mo. 1988.
21. Gastegeir J, Rudolph C, Sadowski J. Automatic Generation of 3D Atomic Coordinates for Organic Molecules. Tetrahedron Computer Methodology. 1990; 3: 537-547.
22. Bajorath J, editor. Chemoinformatics – Concepts, Methods & Tools for Drug Discovery. New Jersey: Humana press Inc.; 2004.
23. Ebe T, Sanderson KA, Wilson PS. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 2. The MARPAT File. J Chem Inf Comput Sci. 199; 31: 31-36.
24. Meehan P, Schofield H. Crossfire: A Structural Revolution for Chemists. Online Information Review. 2001; 25: 241- 249.
25. Blake JE, Dana RC. CASREACT – More than A Million Reactions. J Chem Inf Comput Sci. 1990; 30: 394-399.
26. Wipke WT, Krishnan S, Ouchi GT. Hash Functions for Rapid Storage and Retrieval of Chemical Structure. J Chem Inf Comput Sci.1978; 18: 32 – 37.
27. Ihlenfeldt WD, Gastegeir J. J Computational Chem. 1994; 15: 793-813.
28. Read RC, Corneil DG. The Graph Isomorphism Disease. J Graph Theory. 1977; 1: 339 – 363.
29. Barnard JM. Substructure Searching Methods: Old and New. J Chem Inf Comput Sci. 1993; 33: 532 – 538.
30. Ullmann JR. An Algorithm for Subgraph Isomorphism. J Association for Computing Machinery. 1976; 23: 31 – 42.
31. Johnson MA, Maggiora GM, editors. Concepts and applications of Molecular Similarity. New York: Wiley; 1990.
32. Dean PM, editor. Molecular Similarity in drug design. Glasgow: Chapman and Hall; 1994.
33. Willet P, Barnard JM, Downs GM. Chemical Similarity Searching. J Chem Inf Comput Sci. 1998; 38: 983-996.
34. Dean PM, Lewis RA, editors. Molecular Diversity in drug design. Amsterdam: Kluwer; 1999.
35. Ghose AK, Viswananadhan VN, editors. Combinatorial library design and evaluation: principles, software tools and applications in drug discovery. New York: Marcel Dekker; 2001.
36. Kubinyi H. Similarity and Dissimilarity – a medicinal chemist's view. Perspect. Drug Discov. Design. 1998; 11: 225-252.
37. Martin YC, Kofron JL, Traphagen LM. Do Structurally Similar Molecules Have Similar Biological Activies? J Med Chem. 2002; 45: 4350-4358.
38. Willet P, Winterman V, Bawden D. Implementation of Nearest Neighbour Searching in an Online Chemical Structure Search System. J Chem Inf Comput Sci. 1986; 26: 36-41.
39. Altschul SF, Madden TL, Schaffer AA et.al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997; 25: 3389-3402.
40. Pearson WR. Rapid and Sensitive Sequence Comparison with FASTP and FASTA. Methods Enzymology. 1990; 183: 63-98