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

Values and Practices of Cheminformatics in Chemical World

*Radhika Gupta**
Department of Chemistry, Dr. Hari Singh Gaur Sagar University, Sagar, Madhya Pradesh, India

ABSTRACT

Cheminformatics is the use of computer and informational techniques applied to a range of difficulties in the arena of Chemical Sciences. In silico techniques are used, for example, in pharmaceutical companies in the process of drug discovery. These methods can also be used in chemical and allied industries in various other forms.

Keywords: cheminformatics, drug discovery, pharmaceutical company, value and practices of cheminformatics

Corresponding Author
E-mail: g.rads654@gmail.com

INTRODUCTION

The line "Change is must and change is accelerating" is very important in human life. There are several changes occur in each and every aspects of human civilization from the age of *Homo erectus* to today informational age. The main component of information age is computer which can stored a lot of information giving birth of a discipline namely Informatics. Informatics is Informatics is the discipline of science which investigates the structure and properties (not specific content) of scientific information, as well as the regularities of scientific information activity, its theory, history, methodology and organization. The science of informatics is applied indifferent field of science giving birth of different discipline namely bioinformatics, chemoinformatics, geoinformatics, health informatics, laboratory informatics, neuroinformatics, social informatics [1].

The term "Chemoinformatics" appeared a few years ago and rapidly gained widespread use. Workshops and symposia are organized that are exclusively devoted to chemoinformatics, and many job advertisements can be found in journals. The first mention of chemoinformatics may be attributed to Frank Brown. The use of information technology and management has become a critical part of the drug discovery process as well as to solve the chemical problems. So, 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. Whereas we see here chemoinformatics focused on drug design. Greg Paris came up with a much broader definition chemoinformatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information [1].

History of Chemoinformatics

The first, and still the core, journal for the subject, *the Journal of Chemical Documentation*, started in 1961 (the name Changed to the *Journal of Chemical Information and computer Science* in 1975). Then the first book appeared in 1971 (Lynch, Harrison, Town and Ash,

Computer Handling of Chemical Structure Information). The first international conference on the subject was held in 1973 at Noordwijkerhout and every three years since 1987. The term chemoinformatics was given by Brown in 1998. With all the problems at hand in chemistry, complex relationships, profusion of data, lack of necessary data, quite early on the need was felt in many areas of chemistry to have resort to informatics methods. These various roots of chemoinformatics often go back more than 40 years into the 1960s [2].

Is It Cheminformatics or Chemoinformatics?

The name of our favorite field maybe cheminformatics or chemoinformatics cheminformatics, molecular informatics, chemical informatics, or even chemobioinformatics. All these options have some advantages. By using short cheminformatics you are saving the keyboard of your computer, chemoinformatics sounds nice in sentences like "... our software solution seamlessly integrates chemoinformatics and bioinformatics ...," and the title "Head of chemobioinformatics" on a business card cannot miss the point. Molecular informatics or chemical informatics is less known, but this also means that you are one of the pioneers on the forefront of a new scientific field. But the name of chemoinformatics and cheminformatics are synonymous in use. In the following table frequencies of words cheminformatics and chemoinformatics in web pages are listed, as determined by a popular search engine Google. The ratio characterizes popularity of term cheminformatics over chemoinformatics [3].

Basics of Chemoinformatics

Cheminformatics combines the scientific working fields of chemistry, computer science and information science for example in the areas of topology, chemical graph theory, information retrieval and data mining in the chemical space. Cheminformatics can also be applied to data analysis for various industries like paper and pulp, dyes and such allied industries.

The various fields outlined in the previous section have grown from humble beginnings 40 years ago to areas of intensive activities (Figure 1). On top of that it has been realized that these areas share a large number of common problems, rely on highly related data, and work with similar methods [3]. Thus, these different areas have merged to a discipline of its own:

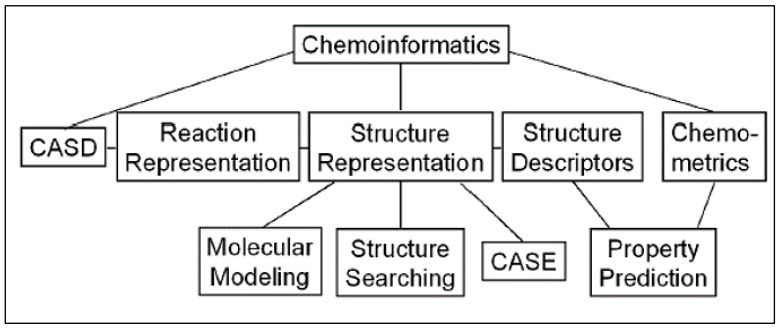

Fig. 1. *The various areas of activities in chemoinformatics.*

Use of Informatics Methods in Chemistry

First of all, chemistry has produced an enormous amount of data and this data avalanche is rapidly increasing. More than 45 million chemical compounds are known and this number is increasing by several millions each year. Novel techniques such

as combinatorial chemistry and high-throughput screening generate huge amounts of data. All this data and information can only be managed and made accessible by storing them in proper databases. That is only possible through chemoinformatics.

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 is where chemoinformatics methods can come in (Figure 2).

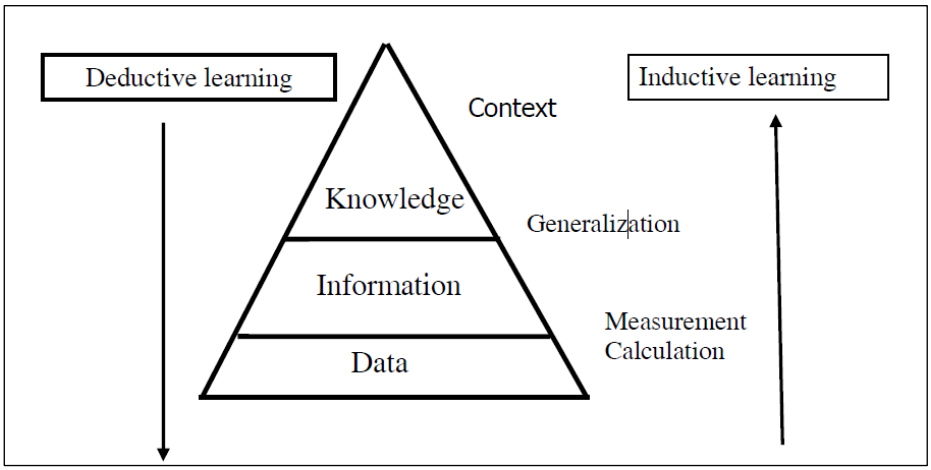

Fig. 2. *Cheminformatics methodology in chemistry.*

Extracting knowledge from chemical information -lots of data (structure, activities, genes, etc.) i.e. called as inductive learning. When we extract data from knowledge, it is called as deductive learning [4].

APPLICATIONS OF CHEMINFORMATICS

Chemical Database

The primary application of cheminformatics is in the storage, indexing and search of information relating to compounds. The efficient search of such stored information includes topics that are dealt with in computer science as data mining, information retrieval, information extraction and machine learning [5].

Related research topics include:

- Unstructured data
- Information retrieval
- Information extraction
- Structured data mining and mining of structured data
- Database mining
- Graph mining
- Molecule mining
- Sequence mining
- Tree mining
- Digital libraries

CHEMICAL FILE FORMAT

The in silico representation of chemical structures uses specialized formats such as the XML-based Chemical Markup Language or SMILES. These representations are often used for storage in large chemical databases. While some formats are suited for visual representations in 2 or 3 dimensions, others are more suited for studying physical interactions, modeling and docking studies [6].

VIRTUAL LIBRARIES

Chemical data can pertain to real or virtual molecules. Virtual libraries of compounds may be generated in various ways to explore chemical space and hypothesize novel compounds with desired properties. Virtual libraries of classes of compounds

(drugs, natural products, diversity-oriented synthetic products) were recently generated using the fragment optimized growth (FOG) algorithm. This was done by using cheminformatic tools to train transition probabilities of a Markov chain on authentic classes of compounds, and then using the Markov chain to generate novel compounds that were similar to the training database [6, 7].

VIRTUAL SCREENING

In contrast to high-throughput screening, virtual screening involves computationally screening in silico libraries of compounds, by means of various methods such as docking, to identify members likely to possess desired properties such as biological activity against a given target. In some cases, combinatorial chemistry is used in the development of the library to increase the efficiency in mining the chemical space. More commonly, a diverse library of small molecules or natural products is screened [7].

Quantitative Structure-Activity Relationship (QSAR)

This is the calculation of quantitative structure-activity relationship and quantitative structure property relationship values, used to predict the activity of compounds from their structures. In this context, there is also a strong relationship to chemometrics (Figure 3). Chemical expert systems are also relevant, since they represent parts of chemical knowledge as an in silico representation. There is a relatively new concept of matched molecular pair analysis or prediction-driven MMPA which is coupled with QSAR model in order to identify activity cliff [7].

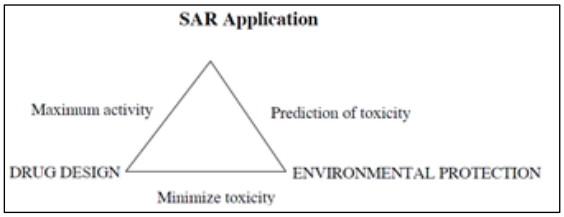

Fig. 3. *Quantitative structure-activity relationship (QSAR).*

Chemical Structure Representation

In the early sixties, various forms of machine readable chemical structure representations were explored as a basis for building databases of chemical structures and reactions. Eventually, connection tables that represent molecules by lists of the atoms and of the bonds in a molecule gained universal acceptance. Connection tables were also used for the Chemical Abstracts Registry System which appeared in the second half of the sixties [8].

A connection table stores the same information that is present in a 2D structure diagram, namely the atoms that are present in a molecule and what bonds exist between the atoms. However, it is stored in a table form which is much easier for a computer to work with. Before a connection table is produced, the atoms in the molecule must be numbered, and an *atom lookup table* produced (Figure 4). This simply stores atom information (usually just the atom type) cross referenced with the atom number. Here is a numbering and atom lookup table for acetaminophen:

Num	Atom Type
1	C
2	C
3	C
4	N
5	C
6	O
7	C
8	C
9	C
10	C
11	O

Fig. 4. *Chemical structure presentation.*

The atom lookup table describes the atoms present in a molecule, but says nothing about how they are connected. The connection table describes how atoms are connected by bonds, and has a row and a column for each atom, the row and column

number representing the number given to the atom [8].

Structure Searching

This involves searching a database for an exact match with a specified query structure. For example, if the following is the query (Figure 5).

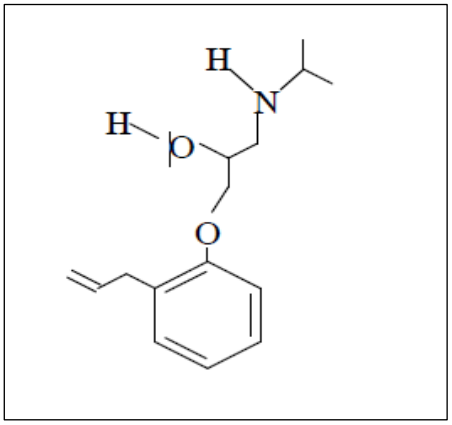

Fig. 5. *Chemical structure query.*

Then only an exact match to this structure would be returned by a search. The techniques used to perform the search would not be covered here, but basically they involve treating the 2D connection table as a mathematical graph, where the nodes represent atoms and the edges represent bonds, and then a test for exact match can be done using a *graph isomorphism* algorithm (a standard computer science technique) [9].

Fingerprint Representations

A fingerprint characterizes the 2D structure of a molecule, usually through a string of '1's and '0's. There are two basic types of fingerprint: structural keys and hashed fingerprints.

Structural Keys

Structural keys contain a string of bits ('1's and '0's) where each bit is set to 1 or 0 depending on the presence or absence of a particular fragment. They usually employ a pre-defined dictionary of fragments.

Hashed Fingerprints

In hashed fingerprints, there is no set dictionary or 1:1 relationship between bits and features. All possible fragments in a compound are generated. The number of fragments represented can be huge. Thus, rather than assigning one bit position for each fragment, the bits are "hashed" down onto a fixed number of bits. Thus, hashed fingerprints are a less precise form, but they carry more information. Once fingerprint representations are available, *similarity coefficients* can be used to give a measure of similarity between two fingerprints [9].

Chemometrics

Initially, the quantitative analysis of chemical data relied exclusively on multilinear regression analysis. However, it was soon recognized in the late sixties that the diversity and complexity of chemical data need a wide range of different and more powerful data analysis methods [10].

Chemical structure representation was introduced in the seventies to analyze chemical data. In the nineties, artificial neural networks gained prominence for analyzing chemical data. The growing of this area led to the establishment of chemometrics as a discipline of its own with its own society, journals, and scientific meetings (Figure 6.).

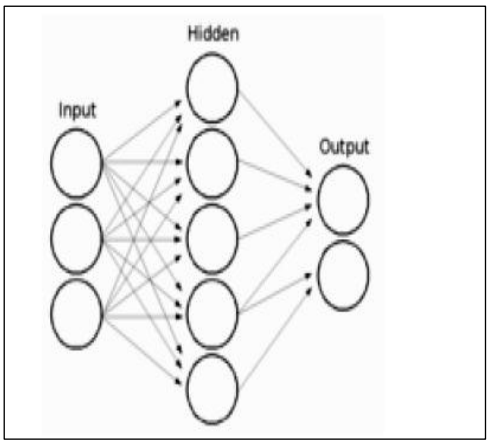

Fig. 6. *Chemical structure representation.*

An artificial neural network (ANN) or commonly just neural network (NN) is an interconnected group of artificial neurons that uses a mathematical model or computational model for information processing based on a connectionist approach to computation [10].

Molecular Modeling

In the late sixties, R. Langridge and coworkers developed methods for visualizing 3D molecular models on the screens of Cathode Ray Tubes. At the same time, G. Marshall started visualizing protein structure on graphic screens. The progress in hardware and software technology, particularly as concerns graphics screens and graphics cards, has led to highly sophisticated systems for the visualization of complex molecular structures in great detail. Programs for 3D structure generation, for protein modeling, and for molecular dynamics calculations have made molecular modeling a widely used technique. The commonly available software for molecular modeling are ArgusLab, Chimera, and Ghemical [10].

COMPUTER-ASSISTED STRUCTURE ELUCIDATION (CASE)

The elucidation of the structure of a chemical compound, be it a reaction product or a compound isolated as a natural product, is one of the fundamental tasks of a chemist. Structure elucidation has to consider a wide variety of different types of information mostly from various spectroscopic methods, and has to consider many structure alternatives. Thus, it is an ambitious and demanding task. It is therefore not surprising that chemists and computer scientists had taken up the challenge and had started in the 1960 ı fs to develop systems for computer-assisted structure elucidation (CASE) as a field of exercise for artificial intelligence techniques. The DENDRAL project, initiated in 1964 at Stanford University gained widespread interest. Other approaches to computer-assisted structure elucidation were initiated in the late sixties by Sasaki at Toyohashi University of Technology and by Munk at the University of Arizona [10].

COMPUTER-ASSISTED SYNTHESIS DESIGN (CASD)

The design of a synthesis for an organic compound needs a lot of knowledge about chemical reactions and on chemical reactivity. Many decisions have to be made between various alternatives as to how to assemble the building blocks of a molecule and which reactions to choose. Therefore, computer-assisted synthesis design (CASD) was seen as a highly interesting challenge and as a field for applying artificial intelligence techniques. In 1969 Corey and Wipke presented their seminal work on the first steps in the development of a synthesis design system. Nearly simultaneously several other groups such as Ugi and coworkers, Hendrickson and Gelernter reported on their work on CASD systems. Later also at Toyohashi work on a CASD system was initiated [11].

Representation of Chemical Reactions

Chemical reactions are represented by the starting materials and products as well as by the reaction conditions. On top of that, one also has to indicate the reaction site, the bonds broken and made in a chemical reaction. Furthermore, the stereochemistry of reactions has to be handled. Searching databases of reactions is a little different to straight searching, although the kinds of search are the same (structure, substructure, similarity). However, searching may be done on reactants, products, or both, and searches may be performed for entire reactions (as opposed to single structures). Representation of reactions is by the usual means (connection tables, atom lookup tables), but with additional information about which molecules are products and reagents, and which reagent atoms map to

which product atoms. A derivative of SMILES, called *Reaction SMILES* is available for representing reactions, along with a way for defining reaction queries called *SMIRKS* [11].

DATA IN CHEMISTRY

Much of our chemical knowledge has been derived from data. Chemistry offers a rich range of data on physical, chemical, and biological properties: binary data for classification, real data for modeling, and spectral data having a high information density. These data have to be brought into a form amenable to easy exchange of information and to data analysis [12].

Datasources and Databases

The enormous amount of data in chemistry has led quite early on to the development of databases to store and disseminate these data in electronic form. Databases have been developed for chemical literature, for chemical compounds, for 3D structures, for reactions, for spectra, etc. The internet is increasingly used to distribute data and information in chemistry. The databases of virtual molecules are available now i.e. the molecules which are not present in the nature, but by just virtually we can prepare databases with the help of databases of other molecules. The commonly available softwares for databases are Amicbase, Asinex Gold, Cheminformatics.org, FDA MRTD, NCI, Otava Dataset, PubChem, and ZINC [12].

Calculation of Structure Descriptors

In most cases, however, physical, chemical, or biological properties cannot be directly calculated from the structure of a compound. In this situation, an indirect approach has to be taken by, first, representing the structure of the compound by structure descriptors, and, then, to establish a relationship between the structure descriptors and the property by analyzing a series of pairs of structure descriptors and associated properties by inductive learning methods. A variety of structure descriptors has been developed encoding 1D, 2D, or 3D structure information or molecular surface properties. The manipulation and analysis of chemical structure information is made through the molecular structure descriptors. These are the numerical values which characterizes properties of molecules. They may represent the physiochemical properties of a molecule or may be the values derived from the algorithm technique to the chemical structures. For example, the molecular weight does not represent the whole properties of a molecule but it is very quick. In case of quantum molecular based structure descriptors, it tells about the properties of a molecule but it is time consuming [13].

Data Analysis Methods

A variety of methods for learning from data, of inductive learning methods is being used in chemistry: statistics, pattern recognition methods, artificial neural networks, genetic algorithms. These methods can be classified into unsupervised and supervised learning methods and are used for classification or quantitative modeling. The softwares are using in data analysis & statistics are ChemTK Lite, PowerMV, and GCluto [13, 14].

Chemistry Based Data Mining and Exploration

For synthesis a molecule, first we have to search data with the help databases available for that molecule, then we have to search the database available for structure analogue. Now the Structure activity relationships are studied and different biological or mechanistic analogue are synthesized (Figure 7). The scheme is given in below.

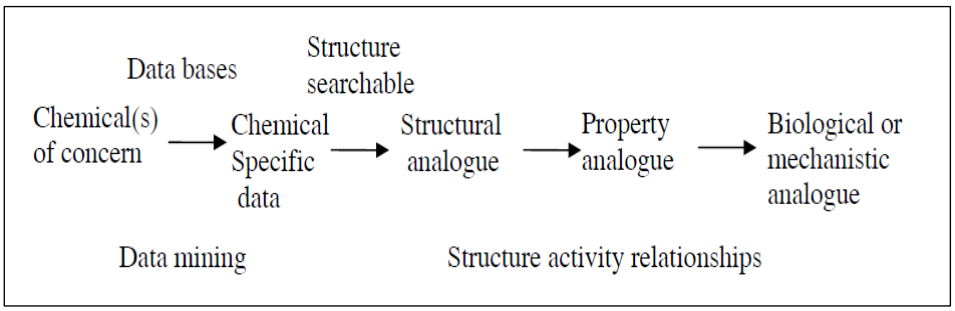

Fig. 7. *Chemistry-based data mining and exploration.*

Chemobioinformatics

Biochemoinformatics (or chemobioinformatics) is a new term to describe the research efforts on meeting the emerging needs for the integration of bioinformatics and chemoinformatics. Historically, bioinformatics and chemoinformatics have largely evolved independently from biology and chemistry. Generally speaking, bioinformatics deals with biological information, which although traditionally refers to sequences information on large biological molecules such as DNA, RNA and proteins, also refers to the more recent emergence of micro array data on gene and protein expression. Chemoinformatics on the other hand mainly deals with chemical information of drug-like small molecules, the molecular weight of these being several hundred Daltons. The elemental data record in bioinformatics is centered on genes and their products (RNA, protein, and so on), whereas the fundamental data type in chemoinformatics is centered on small molecules [14].

CONCLUSIONS

Chemoinformatics has developed over the last 40 years to a mature discipline that has applications in any area of chemistry. Chemoinformatics is the science of determining those important aspects of molecular structures related to desirable properties for some given function. One can contrast the atomic level concerns of drug design where interaction with another molecule is of primary importance with the set of physical attributes related to ADME, for example. In the latter case, interaction with a variety of macromolecules provides a set of molecular filters that can average out specific geometrical details and allows significant models developed by consideration of molecular properties alone. The field has gained so much in importance that the major topics of chemoinformatics have to be integrated into chemistry curricula, a few universities have to offer full chemoinformatics curricula to satisfy the urgent need for chemoinformation specialists. There are still many problems that await a solution and therefore we still will see many new developments in chemoinformatics.

REFERENCES

[1] K. Bhat, C. Bock, N.J. Howard. COS and HTS design of high-performance, nontoxic chemicals for textiles, *NTC Project: C00-PH01*. (formerly C00-P01).

[2] F.K. Brown. Chemoinformatics: what is it and how does it impact? *Drug Discov Ann Rep Med Chem*. 1998; 33: 375–84p.

[3] D.E. Clark, S.D. Pickett. Computational methods for the prediction of ‘drug likeness, *Drug Discov Today*. 2000; 5: 49–58p.

[4] J. Drews. Drug discovery: a historical perspective, *Science*. 2000; 287 5463: 1960–4p.

[5] J. Gasteiger, K. Funatsu. Chemoinformatics – an important scientific discipline, *J Comput Chem Jpn*. 2006; 5(2): 53–8p.

[6] J. Gasteiger (ed.), *Handbook of cheminformatics – From Data to Knowledge*. Weinheim: Wiley-VCH; 2003.
[7] J. Gasteiger, J.T. Engel (eds.), *Chemoinformatics – A Textbook.* Weinheim: Wiley-VCH; 2003.
[8] J. Zupan, J. Gasteiger. *Neural Networks in Chemistry and Drug Design*. 2nd Edn., Weinheim: Wiley-VCH; 1999.
[9] A.R. Leach, V.J. Gillet. *An Introduction to Chemoinformatics*. Springer; 2003,1–57p.
[10] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv Drug Deliv Rev*. 1997; 23: 3–25p.
[11] T.I. Oprea, A.M. Davis, S.J. Teague, P.D. Leeson. Is there a difference between leads and drugs? A historical perspective, *J Chem Inf Comput Sci*. 2001; 41: 1308–15p.
[12] R.K. Lindsay, B.G. Buchanan, E.A. Feigenbaum, J. Lederberg. Applications of artificial intelligence for organic chemistry, *The Dendral Project*. New York: McGraw-Hill; 1980.
[13] J.D. Wild. Getting started in chemoinformatics, Version 1.0, September 2004 Woo, *Environ Carc Ecotox Rev*. 1996; C14: 1–42p.
[14] J. Xu, A. Hagler. Chemoinformatics and drug discovery, *Molecules*. 2002; 7: 566–600p.