

# Chemoinformatics: Principles and Applications

Md. Wasim Aktar<sup>1\*</sup> and Sidhu Murmu<sup>2</sup>

<sup>1</sup>Pesticide Residue Laboratory, Department of Agricultural Chemicals,

<sup>2</sup>Department of Agricultural Chemistry and Soil Science,  
Bidhan Chandra Krishi Viswavidyalaya, Mohanpur-741252,  
Nadia, West Bengal, India

## 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 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.

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. Clearly, the transformation of data into information and of information into knowledge is an endeavor needed in any branch of chemistry not only in drug design. The view that chemoinformatics methods are needed in all areas of chemistry and adhere to a much broader definition: chemoinformatics is the application of informatics methods to solve chemical problems.

---

\*Correspondence to: wasim04101981@yahoo.co.in

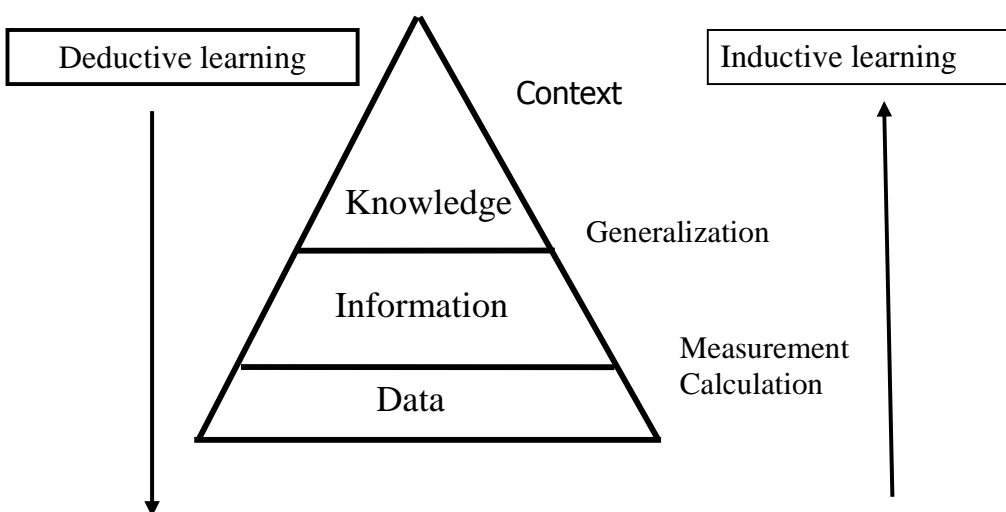
## Why do we have to use informatics methods in chemistry?

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.

On the other hand, for many problems the necessary information is not available. We know the 3D structure, determined by X ray crystallography for about 300,000 organic compounds. Or, as another point, the largest database of infrared spectra contains about 200,000 spectra. Although these numbers may seem large, they are small in comparison to the number of known compounds: We know from less than 1% of all compounds their 3D structure or have their infrared spectra. 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? There is another reason why we need informatics methods in chemistry: Many problems in chemistry are too complex to be solved by methods based on first principles 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 is where chemoinformatics methods can come in.

### The representation of the chemoinformatics in graphical form is given below:



Source: authors

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.

### **Is it Cheminformatics or Chemoinformatics?**

The name of our favourite field maybe cheminformatics or chemoinformatics chemiinformatics, 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.

Year	Cheminformatics	Chemoinformatics	Ratio
2000	39	684	0.05
2001	8,010	2,910	2.75
2002	34,000	16,000	2.12
2203	58,143	32,872	1.77
2204	85,435	60,439	1.41
2005	6,58,298	2,72,096	2.41
2006	3,17,000+	1,63,000+	1.94

*Source:* Leach AR. *et.al.* (2003)

### **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.

## 1. 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. 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. 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

*Source:* authors

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.

	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>	<b>11</b>
<b>1</b>	0	<b>1</b>	0	0	0	0	0	0	0	<b>2</b>	0
<b>2</b>	<b>1</b>	0	<b>2</b>	0	0	0	0	0	0	0	0
<b>3</b>	0	<b>2</b>	0	<b>1</b>	0	0	0	<b>1</b>	0	0	0
<b>4</b>	0	0	<b>1</b>	0	<b>1</b>	0	0	0	0	0	0
<b>5</b>	0	0	0	<b>1</b>	0	<b>2</b>	<b>1</b>	0	0	0	0
<b>6</b>	0	0	0	0	<b>2</b>	0	0	0	0	0	0
<b>7</b>	0	0	0	0	<b>1</b>	0	0	0	0	0	0
<b>8</b>	0	0	<b>1</b>	0	0	0	0	0	<b>2</b>	0	0
<b>9</b>	0	0	0	0	0	0	0	<b>2</b>	0	<b>1</b>	0
<b>10</b>	<b>2</b>	0	0	0	0	0	0	0	<b>1</b>	0	<b>1</b>
<b>11</b>	0	0	0	0	0	0	0	0	0	<b>1</b>	0

*Source:* authors

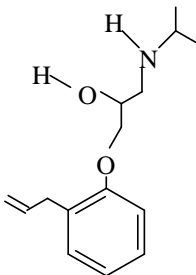
For example, if a bond exists between atom 5 and atom 8, then a “1” is placed at the intersection of row 5 and column 8 (and also row 8 and column 5), otherwise a 0 is placed at the intersection. Further, we may use a 2 to represent a double bond, 3 to represent a triple bond, and so on. Here is the connection table for Acetaminophen, along with the diagram showing which numbers correspond to which atoms. For clarity, the non-zero entries are showing in bold. Note how the table is symmetrical about the diagonal from top left to bottom right. This will always be the case since, for example, if atom 3 is bonded to atom 2, then atom 2 is also by definition bonded to atom 3. Since this connection table effectively stores each piece of information twice, it is called a redundant connection *table*. Normally, we just store one half of the table in a non-redundant connection table as shown below:

	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>9</b>	<b>10</b>	<b>11</b>	<b>12</b>
<b>1</b>											
<b>2</b>	<b>1</b>										
<b>3</b>	0	<b>2</b>									
<b>4</b>	0	0	<b>1</b>								
<b>5</b>	0	0	0	<b>1</b>							
<b>6</b>	0	0	0	0	<b>2</b>						
<b>7</b>	0	0	0	0	<b>1</b>	0					
<b>9</b>	0	0	<b>1</b>	0	0	0	0				
<b>10</b>	0	0	0	0	0	0	0	<b>2</b>			
<b>11</b>	<b>2</b>	0	0	0	0	0	0	0	<b>1</b>		
<b>12</b>	0	0	0	0	0	0	0	0	0	<b>1</b>	

*Source:* authors

## 2. Structure Searching

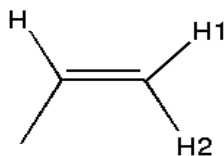
This involves searching a database for an exact match with a specified query structure. For example, if the following is the query.



Then only an exact match to this structure would be returned by a search. The techniques used to perform the search won't 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).

A connection table is essentially a representation of the molecular graph (A graph is a mathematical conceptualization of anything that consists of connected points). Therefore, for storing a unique representation of a molecule and for allowing its retrieval, the graph isomorphism problem had to be solved to define from a set of potential representations of a molecule a single one as the unique one.

The first solution was the Morgan algorithm for numbering the atoms of a molecule in a unique and unambiguous manner. By Morgan algorithm atoms of the same elemental type can be topologically equivalent or not is judged. Let us label the carbons C, C<sub>H</sub> and C<sub>H<sub>1</sub>H<sub>2</sub></sub>, and the hydrogens H, H<sub>1</sub> and H<sub>2</sub>. Obviously, only atoms of the same elemental type can be topologically equivalent. Thus, it is immediately clear that the carbon atoms can be separated from the hydrogen atoms.

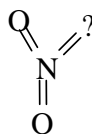


The algorithm proceeds by analyzing the extended connectivity in the following way. A score is assigned to each atom. Initially, the scores are computed by counting the number of bonds formed by each atom: *i.e.* C = 1, C<sub>H</sub> = 3 and C<sub>H<sub>1</sub>H<sub>2</sub></sub> = 3. This tells us that C is unique; hence, amongst the carbons, only C<sub>H</sub> and C<sub>H<sub>1</sub>H<sub>2</sub></sub> can possibly be topologically equivalent. All the hydrogens have a score (*i.e.* sum connectivity) of 1. In the second

iteration, the new score of each atom is calculated by summing the first-iteration scores of all the atoms to which it is bonded.  $C_H$  gets a score of  $1 (C) + 1 (H) + 3 (C_{H_1H_2}) = 5$ .  $C_{H_1H_2}$  gets a score of  $3 (C_H) + 1 (H_1) + 1 (H_2) = 5$ . H gets a score of 3.  $H_1$  and  $H_2$  also get scores of 3. Scores based on summing the atomic numbers of bound atoms are also computed:  $C_H$  gets a score of 13,  $C_{H_1H_2}$  gets a score of 8 and the protons all score 6. This means that  $C_H$  is distinct from  $C_{H_1H_2}$ . In the third cycle of iteration, the scores based on numbers of bonds become 5 for all the protons, but the scores based on atomic numbers become 13 for H, and 8 for  $H_1$  and  $H_2$ . Thus, H is distinct from  $H_1$  and  $H_2$ . The termination criterion for the iterative process is when no further atoms can be assigned as unique by an iteration. At this point, we know which atoms are grouped together: those that had the same score at each iteration are topologically equivalent. In this example, the fourth pass shows that  $H_1$  and  $H_2$  are equivalent. This provided the basis for full structure searching. Then, methods were developed for substructure searching, for similarity searching, and for 3D structure searching.

### Substructure searching

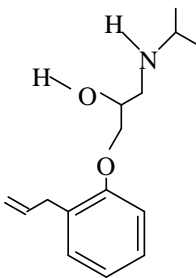
A substructure search involves finding all the structures in a database that contain one or more particular structural fragments. For example, we might want to find all of the structures in a database which contain the nitro group:



Substructure searching requires some method of specifying a query (i.e., we want to find *this* and *that*, but not *this*, etc). One popular example is SMARTS, an extension to SMILES. Mathematically, substructure searching is performed, as with structure searching, using a graph representation, but this time a *subgraph isomorphism algorithm* finds occurrences of subgraphs (i.e. substructures) in a structure.

### Similarity searching

Similarity searching involves looking for all the structures in a database that are highly similar to a given structure. The most common use is to find compounds that could exhibit similar properties (based on the similar property principle that compounds with similar structures are likely to exhibit similar biological behaviors). Note that “similarity” is a subjective thing. As an example, a similarity search might involve looking for structures with a similarity greater than 0.7 to this molecule



Obviously some method is required for measuring similarity. This is usually done using fingerprint representations and similarity coefficients as described below, which are used in various applications that involve measurement of similarity, for example cluster analysis.

### 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.

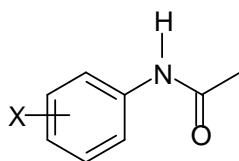
Once fingerprint representations are available, *similarity coefficients* can be used to give a measure of similarity between two fingerprints.

### 3. Quantitative Structure Activity / Property Relationship (QSAR/QSPR)

Building on work by Hammett and Taft in the fifties, Hansch and Fujita showed in 1964 that the influence of substituents on biological activity data can be quantified. In the last 40 years, an enormous amount of work on relating descriptors derived from molecular structures with a variety of physical, chemical, or biological data has appeared. These studies have established Quantitative Structure-Activity Relationships (QSAR) and Quantitative Structure-Property Relationships (QSPR) as fields of their own, with their own journals, societies, and conferences.

**Percent Spikelet Sterility (% Ss) of N-acylanilines Tested in Winter 2001-02 at 1500 ppm Spray Concentrations on PBW 343**



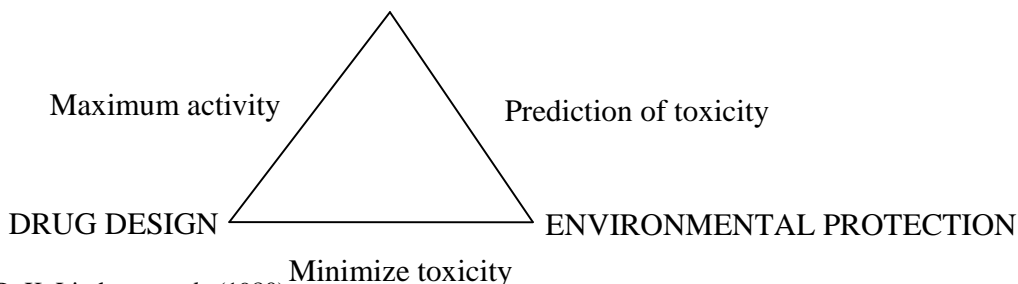


Ethyl Oxanilates (R= COOEt)			Elkyl Oxanilates			
No.	X	Ss (%)	No.	X	R	Ss (%)
1	H	64.18	28	4-F	COOMe	64.32
2	2-F	68.13	29	4-F	COOiPr	67.15
3	3-F	50.04	30	4-F	COCH <sub>3</sub>	51.71
4	3-Cl, 4-F	77.01	31	4-F	COOC <sub>2</sub> H <sub>4</sub> OMe	67.07
5	4-F	69.97	32	4-F	CH <sub>2</sub> COOEt	64.66
6	4- Br	69.06	33	4-Br	CH <sub>2</sub> COOEt	63.40
7	2-Cl	50.25	34	H	CH <sub>2</sub> COOEt	62.07
8	3-Cl	44.25	35	2-OMe	CH <sub>2</sub> COOEt	25.51
9	2,4-Cl <sub>2</sub>	41.56	36	3-OMe	CH <sub>2</sub> COOEt	20.54
10	4-Cl	72.09	37	2-NO <sub>2</sub>	CH <sub>2</sub> COOEt	12.53
11	2-OMe	78.02	38	4-F	CH <sub>2</sub> COOCH <sub>3</sub>	69.12
12	3-OMe	39.07	39	4-Br	CH <sub>2</sub> COOCH <sub>3</sub>	63.65
13	4-OMe	64.39	40	H	CH <sub>2</sub> COOCH <sub>3</sub>	64.52
14	2,4-(OMe) <sub>2</sub>	63.43	41	2-Cl	CH <sub>2</sub> COOCH <sub>3</sub>	34.03
15	2-NO <sub>2</sub>	61.50	42	3-Cl	CH <sub>2</sub> COOCH <sub>3</sub>	15.53
16	3-NO <sub>2</sub>	32.13	43	4-Cl	CH <sub>2</sub> COOCH <sub>3</sub>	58.65
17	4-NO <sub>2</sub>	79.06	44	3-NO <sub>2</sub>	CH <sub>2</sub> COOCH <sub>3</sub>	22.13
18	2,4-(NO <sub>2</sub> ) <sub>2</sub>	62.37	45	3-CH <sub>3</sub>	CH <sub>2</sub> COOCH <sub>3</sub>	2.04
19	3-Me	23.29	46	4-F	CH <sub>3</sub>	61.35
20	2-CN	61.43	47	4-F	CH <sub>2</sub> Cl	39.02
21	3-CN	71.74	48	4-Br	CH <sub>2</sub> Cl	39.45
22	4-CN	66.63	49	H	CH <sub>2</sub> Cl	32.16
23	2-CF <sub>3</sub>	64.59	50	4-F	CHCl <sub>2</sub>	63.75
24	3-CF <sub>3</sub>	64.86	51	4-Br	CHCl <sub>2</sub>	62.39
25	4-CF <sub>3</sub>	69.57	52	H	CHCl <sub>2</sub>	34.85
26	4-Et	18.49	53	4-F	CCl <sub>3</sub>	69.61
27	4-iPr	9.41	54	4-Br	CCl <sub>3</sub>	68.69
			55	H	CCl <sub>3</sub>	46.07
Emulsion Control		0.46				
CD ( p= 0.05)		0.59				

Source: Gasteiger J. *et.al.* (2006)

Modern QSAR involves applying artificial intelligence and Statistical techniques to 2D or 3D molecular representations.

### SAR Application



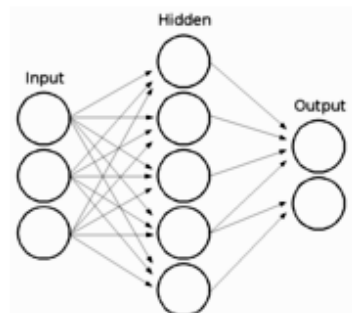
Source: R. K. Lindsay *et. al.* (1980).

At the time of drug design, we have to look after these following points-

- Single therapeutic target
- Drug like chemical
- Some toxicity anticipated
- Multiple unknown targets
- Diverse Structures
- Human and ecosystems

#### 4. 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. Pattern recognition methods were 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.



Source: R. K. Lindsay *et. al.* (1980).

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.

## **5. 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 softwares for molecular modeling are ArgusLab, Chimera, and Ghemical.

## **6. 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 1960s 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.

## **7. 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.

## **Basics of Chemoinformatics**

The various fields outlined in the previous section have grown from humble beginnings 40 years ago to areas of intensive activities. 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. Thus, these different areas have merged to a discipline of its own: Chemoinformatics.

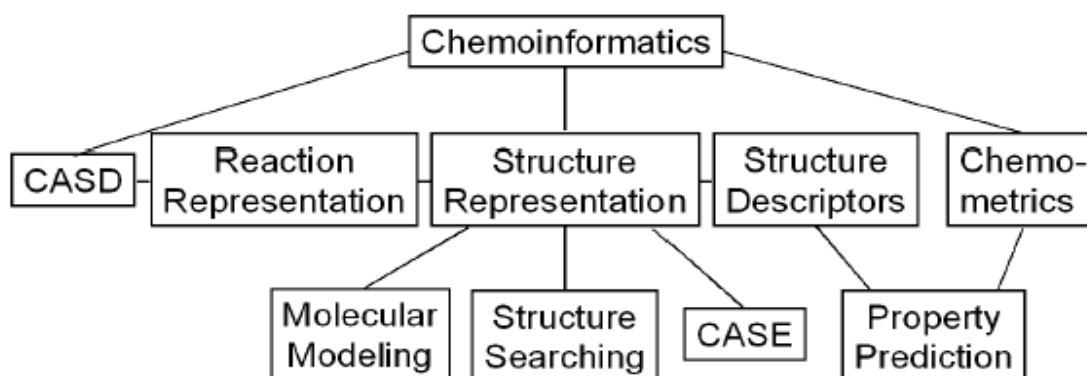


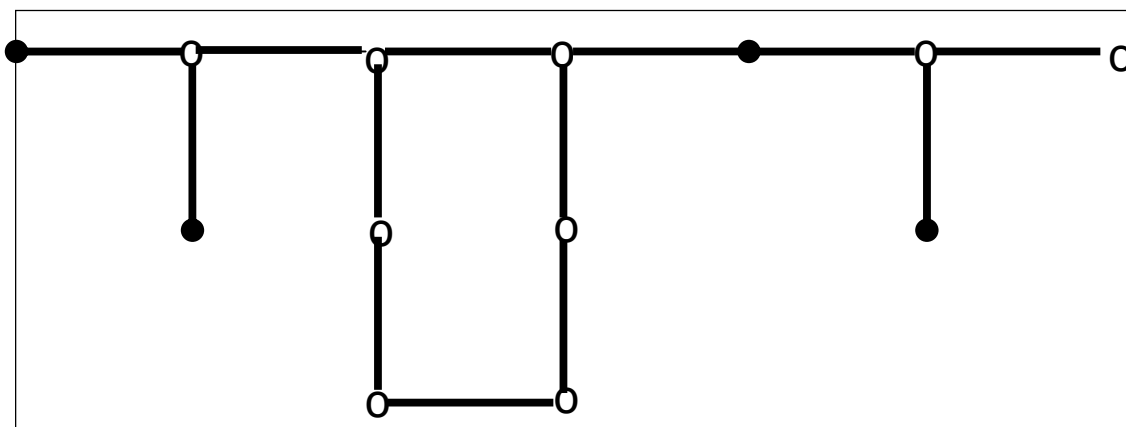
Figure 1. The various areas of activities in chemoinformatics

**Source:** Lipinski, C.A *et.al.*, (1997)

The extent of this field has recently been documented by a "Handbook of Chemoinformatics", covering 73 contributions by 65 scientists on 1850 pages in four volumes. The following gives an overview of chemoinformatics, emphasizing the problems and solutions - common to the various more specialized subfields.

### **1. Representation of Chemical Compounds**

A whole range of methods for the computer representation of chemical compounds and structures has been developed: linear codes, connection tables, matrices. Special methods had to be devised to uniquely represent a chemical structure, to perceive features such as rings and aromaticity, and to treat stereochemistry, 3D structures, or molecular surfaces. Earlier the chemical 2D structure representations are done by software namely Chemdraw, ISIS etc. But now, chemical structures are represented by molecular graph. A graph is an abstract structure that contains nodes connected by edges. Here nodes are represented by atoms and edges by bonds. A graph represents only topology of a molecules i.e. the ways the nodes i.e. atoms are connected.



Aspirin

**Source:** J. Zupan *et.al.*,(1999).

The aspirin structure can be represented by Graph theory, where Oxygen atom is represented by filled bullet and carbon atom is represented by vacant bullet and hydrogen atom is not represented here. So, the aspirin structure will be-

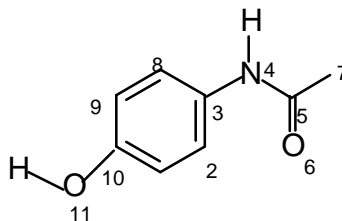
For similarities searching we can use the graph isomorphism or by any algorithm.

### Linear notations

Structure linear notations convert chemical structure connection tables to a string, a sequence of letters, using a set of rules. The earliest structure linear notation was the Wiswesser Line Notation (WLN). ISI® adopted WLN to be used in some of their products in 1968 and, it is still use today. It was also adopted in the mid 1960s for internal use by many pharmaceutical companies. At that time (mid 60s to 80s), it was considered the best tool to represent, retrieve and print chemical structures. In WLN, letters represents structural fragments and a complete structure is represented as a string. This system efficiently compressed structural data and, was very useful to storing and searching chemical structures in low performance computer systems. However, the WLN is difficult for non- experts to understand. Later, David Weininger suggested a new linear notation designated as SMILESTM. Since SMILESTM is very close to the “natural language” used by organic chemists, SMILESTM is widely accepted and used in many chemical database systems. To successfully represent a structure, a linear notation should be canonicalized. That is, one structure should not correspond to more than one linear notation string, and conversely, one linear notation string should only be interpreted as one structure.

Attempt to condense all of the connectivity information into a single text string. The two most popular formats are SMILES (from Daylight) and SLN (Tripos format inspired by SMILES).

## SMILES (Simplified Molecular Input Line Entry Specification)



Acetaminophen

In SMILES, atoms are generally represented by their chemical symbol, with upper-case representing an aliphatic atom (C = aliphatic carbon, N = aliphatic nitrogen, etc) and lower-case representing an aromatic atom (c = aromatic carbon, etc). Hydrogens are not normally represented explicitly. Consecutive characters represent atoms bonded together with a single bond. Therefore, the SMILES for propane would simply be: CCC or 1-propanol would be: CCCO. Double bonds are represented by an “=” sign, e.g. propene would be: C=CC. Parentheses are used to represent branching in the molecule, e.g. the SMILES for Isopropyl alcohol (2-propanol) is: CC(O)C. Atoms other than the major organic ones (C, S, N, O, P, Cl, Br, I, B) or ions must be enclosed in square brackets. Ring enclosures are represented by using numbers to signify attachment points, usually starting at 1. The first occurrence of the number defines the attachment point, and subsequent occurrences indicate that the structure joins back to the attachment point at that position. For example, the SMILES for Benzene is as follows (note the small ‘c’ for aromatic carbon): c1ccccc1. We can also use branching from the ring system, e.g. c1cc(Br)ccc1 represents bromobenzene. Note that in many cases there can be several SMILES to represent the same structure – for example, we could alternatively represent bromobenzene as: c1cccc(Br)c1. So here is a SMILES representation for acetaminophen, the structure at the top of this document: c1c(O)ccc(NC(=O)C)c1. The great advantage of these methods is brevity – for example an entire SMILES string can be stored in a single spreadsheet cell. However, it is hard to add additional information (coordinates, properties, etc) in these formats in an elegant way.

## Canonicalization


If a structure corresponds to a unique WLN or a unique SMILESTM string, then the structure search results in a string match. WLN could meet this requirement in most cases. The SMILESTM approach can do this after canonical processing. Therefore, both WLN and canonical SMILESTM are able to solve structure search problems by string matches. A molecular graph (2D structure) can also be canonicalized into a real number through a mathematical algorithm. The real number is identified as a molecular topologic index. However, two different structures can have the same topologic index. Therefore, topologic indices can only be used as screens for accelerating structure database searching. Actually, the concept of molecular index was originally proposed for QSAR and QSPR studies. Wiener reported the first molecular topological index in 1947 (Brown, 1998). If a molecule and its specific topologic index had a one-to-one relationship, then structure search could be done by number comparison (Brown, 1998). However, substructure search still had to use an atom-by-atom matching algorithm, which, as

mentioned earlier, could be very time-consuming. In order to further enhance chemical database search performance, efforts have been on the way to seek better structural screening technologies.

### Sources of 3d informations and the Representation of molecules in 3D Form.

3D information can be obtained through X-ray crystallography, NMR spectroscopy or by computational means. The basic forms of 3D representation are the *coordinate table* and the *distance matrix*.

A coordinate table is simply an extension of the atom lookup table that also contains coordinates for each atom. These coordinates are relative to a consistent origin. Here is a sample coordinate table for Aspirin, along with a 3D structure with the atoms numbered:



Atom	Label	X	Y	Z
1	C	-1.8920	-0.9920	-1.5760
2	C	-1.3680	-2.1480	-0.9880
3	C	-0.0760	-2.1440	-0.4640
4	C	0.7080	-0.9840	-0.5200
5	C	0.2000	-0.1560	-1.1960
6	C	-0.1080	0.1600	-1.6520
7	O	2.0840	-1.0280	0.1040
8	O	2.5320	-2.0320	0.6360
9	C	2.8760	0.0240	0.1120
10	O	0.7520	1.3320	-1.0840
11	O	0.6680	2.0240	0.0320
12	C	1.3000	3.0600	0.1520
13	C	-0.2400	1.5760	1.4440

**Source:** Gasteiger, J., (2003)

Distance matrices are similar to connection tables, except that instead of storing connectivity information, they store relative distances (in Angstroms) between all atoms. Here is a sample distance matrix for the Aspirin molecule above. Many pattern recognition techniques require distance or similarity measurements to quantitatively measure the distance or similarity of two objects (in our case, the objects are small molecules). Euclidean distance, Mahalanobis distance and correlation coefficients are commonly used for distance measurement,

$$D(\vec{A}, \vec{B}) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2} \quad (1)$$

$$D(\vec{A}, \vec{B}) = \sqrt{(a_i - b_i) \Sigma^{-1} (a_i - b_i)^T} \quad (2)$$

$$R(\vec{A}, \vec{B}) = \frac{\sum_{i=1}^n a_i b_i}{\sqrt{\sum_{i=1}^n a_i^2 \sum_{i=1}^n b_i^2}} \quad (3)$$

**Source:** Clark and Pickett, (2000)

where  $n$  is the number of descriptors,  $D$  represents the absolute distance between  $A$  and  $B$ ,  $R$  represents the angle of vectors  $A$  and  $B$  in multidimensional space and, is interpreted as the quantity of the linear correlation of  $A$  and  $B$ . The value range of  $R$  is between  $-1$  to  $+1$  that is, from 100% dissimilar to 100% similar. The Euclidian distance assumes that variables are uncorrelated. When variables are correlated, the simple Euclidean distance is not an appropriate measure, however, the Mahalanobis distance (2) will adequately account such correlations. The Tanimoto coefficient is commonly employed for similarity measurements of bit-strings of structural fingerprints (Boolean logic). The simplified form is

$$T(A, B) = \frac{\gamma}{\alpha + \beta - \gamma} \quad (4)$$

**Source:** Clark and Pickett, (2000)

where  $\alpha$  is the count of substructures in structure  $A$ ,  $\beta$  the count of substructures in structure  $B$ , and  $\gamma$  is the count of substructures in both  $A$  and  $B$ . Many different similarity calculations have been reported. Holliday, Hu and Willett have published a comparison of 22 similarity coefficients for the calculation of inter-molecular similarity and dissimilarity, using 2D fragment bit-strings (Clark and Pickett, 2000).



	1	2	3	4	5	6	7	8	9	10	11	12	13
1		1.4	2.4	2.8	2.4	3.8	4.8	4.2	1.4	2.4	2.7	2.9	4.3
2			1.4	2.4	2.8	4.3	5.1	5.0	2.4	3.7	3.9	4.2	5.6
3				1.4	2.4	3.8	4.2	4.8	2.8	4.2	4.7	4.9	6.4
4					1.4	2.5	2.8	3.6	2.4	3.7	4.7	4.6	6.1
5						1.5	2.4	2.3	1.4	2.3	3.7	3.5	4.8
6							1.3	1.2	2.5	2.8	4.4	3.9	5.0
7								2.2	3.7	4.1	5.7	5.2	6.3
8									2.8	2.5	4.2	3.5	4.3
9										1.4	2.6	2.3	3.7
10											2.2	1.3	2.5
11												1.2	2.4
12													1.5
13													

Source: Gasteiger, J., (2003)

Distance matrices are useful when comparing molecules with each other, whereas coordinate tables tend to be used for structure visualization.

## 2. 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*.

## 3. 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

#### **4. 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.

#### **5. Structure Search Methods**

In order to retrieve data and information from databases, access has to be provided to chemical structure information. Methods have been developed for full structure, for substructure, and for similarity searching. Those are discussed in above.

#### **6. Methods for Calculating Physical and Chemical Data**

A variety of physical and chemical data of compounds can directly be calculated by a range of methods. Foremost are quantum mechanical calculations of various degrees of sophistication. However, simple methods such as additive schemes can also be used to estimate a variety of data with reasonable accuracy.

#### **7. 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 propertities of molecules. They may represents the physiochemical properties of a molecule or may b 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.

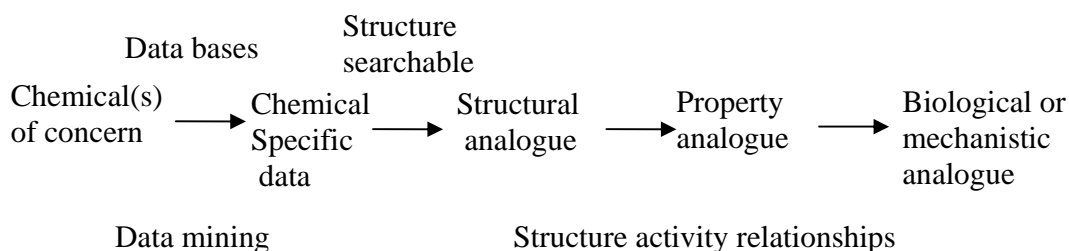
The commonly used molecular descriptors are logP and molar refractivity. Hydrophobicity is most commonly modeled using the logarithm values of partition coefficient i.e. logP.

## 8. 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, & GCluto.

### 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. The scheme is given in below.....



## Applications of Chemoinformatics

### a. Fields of Chemistry

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!

1. Chemical Information
  - storage and retrieval of chemical structures and associated data to manage the flood of data by the softwares are available for drawing and databases.
  - dissemination of data on the internet
  - cross-linking of data to information
2. All fields of chemistry
  - prediction of the physical, chemical, or biological properties of compounds
3. 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
- 4. Organic Chemistry
  - prediction of the course and products of organic reactions
  - design of organic syntheses
- 5. Drug Design as well as for bioactive molecules.
  - 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

Finally, small molecules can be used for docking and drug screening/discovery. Small molecules, as well as their synthetic derivatives, can be docked to a protein target and computationally filtered (e.g. by solubility) to produce a ranked list of candidates that can then be tested in the laboratory. Known ligands can also be used in similarity searches, or as scaffold for further molecular engineering. We will present several recent drug discovery efforts that leverage ChemDB and the computational tools described above. In particular, the discovery of several compounds has done that can bind to the Carboxyltransferase domain of Acyl-CoA Carboxylase, AccD5 from *Mycobacterium tuberculosis*, a new TB therapeutic target.

- prediction of the metabolism of xenobiotics
- analysis of biochemical pathways
- Modeling of ADME-Tox properties.

Historically, drug absorption, distribution, metabolism, excretion, and toxicity (ADMET) studies in animal models were performed after a lead compound was identified. Now, pharmaceutical companies are employing higher-throughput, in vitro assays to evaluate the ADMET characteristics of potential leads at earlier stages of development. This is done in order to eliminate candidates as early as possible, thus avoiding costs, which would have been expended on chemical synthesis and biological testing. Scientists are developing computational methods to select only compounds with reasonable ADMET properties for screening. Molecules from these computationally screened virtual libraries can then be synthesized for high-throughput biological activity screening. As the predictive ability of ADME/Tox software improves, and as pharmaceutical companies incorporate computational prediction methods into their R&D programs, the drug discovery process will move from a screening based to a knowledge-based paradigm. Under multi-parametric optimization drug discovery strategies, there is no excuse for failing to know the relative solubility and permeability rankings of collections of chemical compounds for lead identification.

**a. Absorption.** Passive intestinal absorption (PIA) models have been studied by many groups, for years. The fluid mosaic model holds that the structure of a cell membrane is an interrupted phospholipid bilayer capable of both hydrophilic and hydrophobic interactions. Trans cellular passage through the membrane lipid/aqueous environment is the predominant pathway for passive absorption of lipophilic compounds, while low-molecular-weight (<200), hydrophilic compounds make use of the water-filled channels of the tight junctions between membrane cells (paracellular transport). Therefore, lipophilicity is considered a key property for activity in drug design and is a common property used to estimate the membrane permeability of a molecule. Lipophilicity is measured as the log of the partition coefficient between n-octanol and water (logP). LogP prediction programs are available and results are reasonably good. But, the relationship between logP and permeability is not linear. Permeability drops at both low and high logP. It is theorized that These non-linearities due to: (1) the inability of weakly lipophilic compounds to penetrate the lipid portion of the membrane and (2) the excessive partitioning of strongly lipophilic compounds into the lipid portion of the membrane and their subsequent inability to pass through the aqueous portion of the membrane. A strong relationship between PIA and polar surface area (PSA) has been discovered by several groups. However, the models usually do not take the effects of other descriptors into account. In addition, the datasets used to build the PSA models are small. Even though a wide range of PSA was covered, it is not necessarily true the models cover the entire chemical space. Therefore, linear and non-linear multivariate models have been introduced to model PIA based upon: logP, molecular weight, Hbonding, free energy, H-bond donor, H-bond acceptor, polarizability, numbers and strengths of Hbond acceptor nitrogen and oxygen atoms, number of H-bond donor atoms, and lipophilicity (log D at pH 7.4) on the Caco-2 cell permeability. To select the best descriptors for predictive models, a genetic algorithm has been used.

**b. Distribution.** CNS-active drugs (CNS, central nervous system) must cross the blood-brain barrier (BBB). The experimental determination of the brain-blood partition ratio is difficult and timeconsuming to compute since it involves the direct measurement of the drug concentration in the brain and blood of laboratory animals. This obviously requires the synthesis of the compounds, often in radio labeled form. *In vitro* techniques to predict brain penetration are available, but they are experimentally cumbersome. The earlier work involved in correlating log(C<sub>brain</sub>/C<sub>blood</sub>) or logBB and logP (octanol-cyclohexane), Pyclohexane, or logPoct was based upon smaller (about 20 compounds) data sets. More descriptors have been correlated with logBB, such as: excess molar refraction, solute polarizability, hydrogen bond acidity and basicity, and molecular volume. More recently a regression study on logBB and free energy  $\Delta G$  has been reported. Descriptors derived from 3D molecular fields to estimate the BBB permeation on a larger set of compounds and to produce a simple mathematical model have been studied. The method used (VolSurf) transforms 3D fields into descriptors and correlates them to the experimental permeation by a discriminates partial least squares procedure. Human serum albumin (HSA) protein is the major transporter of non-esterified fatty acids, as well as of different drugs and metabolites, to different tissues. HSA allows solubilization of hydrophobic compounds, contributes to a more homogeneous distribution of drugs in the body, and increases their biological lifetime. The binding

strength of any drug to serum albumin is the main factor for availability of that drug to diffuse from the circulatory system to target tissues. All these factors cause the pharmacokinetics of almost any drug to be influenced and controlled by its binding to serum albumin. Therefore, QSAR study on binding of drugs and metabolites to HSA is extremely important for the drug distribution. Biosensor analysis for prediction of HSA has been reported. In order to build an *in silico* predictive model for binding affinities to HSA, Colmenarejo and coworkers at GlaxoSmithKline used a genetic algorithm to exhaustively search and select for multivariate and non-linear equations, starting from a large pool of molecular descriptors. They found that hydrophobicity (as measured by the ClogP) is the most important variable for determining the binding extent to HSA. Binding to HSA turns out to be determined by a combination of hydrophobic forces together with some modulating shape factors. This agrees with X-ray structures of HSA alone or, bound to ligands, where the binding pockets of both sites I and II are composed mainly of hydrophobic residues.

**c. Metabolism.** Drug metabolism is another barrier to overcome. Metabolism is studied, by *in vitro*, *in vivo* and *in silico* approaches. HTS has been used for metabolism and pharmacokinetics. *In vitro* approaches determine metabolic stability, screening for inhibitors of specific cytochrome P450 isozymes and, identifying the most important metabolites. *In vivo* approaches measure hepatic metabolic clearance, volume of distribution, bioavailability, and, identify major metabolites. *In silico* approaches are categorized into three classes: QSAR and pharmacophore models, protein models, and expert systems. QSAR and pharmacophore models predict substrates and inhibitors of a specific cytochrome P450 isozyme. Protein models rationalize metabolite formations and identify possible substrates, potential metabolites or, inhibitors by means of docking algorithms. Stereoelectronic factors involved in metabolic transformations can be taken into account using quantum chemical calculations. Expert systems are predictive databases that attempt to identify potential metabolites of a compound as determined by knowledge based rules defining the most likely products. Testa advised that in structure-metabolism relationship (SMR) studies, the greater the chemical diversity of the investigated compounds, the smaller the chance that SMRs exist and can be uncovered. On the other hand, the information content of an SMR (if it exists) will increase as the boundaries of the chemical space increases and as the diversity of the compounds under investigation increases. This paradox may limit the capacity of SMR, no matter which approach is used. Keseru and Molnar think efficient PK optimization requires metabolic diversity within the focused library that cannot be achieved by the application of a simple SMR with limited information content. The high degree of structural similarity (especially in combinatorial libraries with a common core) prevents the application in metabolic diversity analysis. Therefore, they introduced a metabolic fingerprint concept, METAPRINT, for the assessment of metabolic similarity and diversity in combinatorial chemical libraries. Their metabolic fingerprint was developed by predicting metabolic pathways and corresponding potential metabolites.

**d. Excretion/Elimination.** Drugs such as the non-steroidal anti-inflammatory drugs (NSAIDs), are used in long term treatment. The accumulation of these drugs in the body may lead to serious side effects. Therefore, the prediction of half-life, which determines

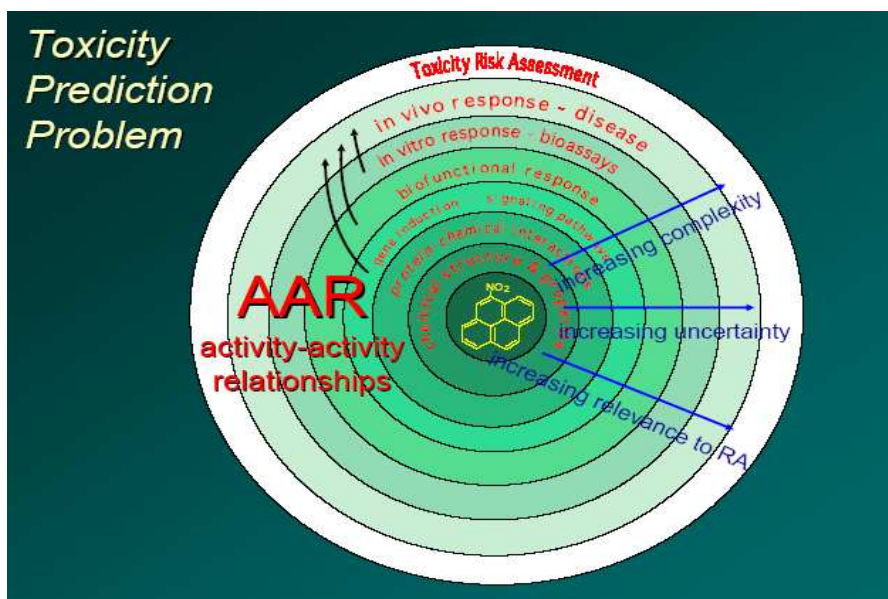
the length of time a drug will persist in the body, is important in order to reduce subsequent drug failures. Prediction of half-life is difficult, due to the multi-faceted nature of drug elimination. Distribution of drug in fat and major organs, excretion by kidneys and metabolism by liver all contribute to the rate at which a drug is eliminated from the body. On the other hand, it may be possible to make use of qualitative predictions of half-life. Such information can be used, for example, to predict whether a drug is likely to accumulate to a significant extent when used for prolonged treatment.

**e. Toxicity.** Many drugs are withdrawn for safety reasons and there are many reasons, including metabolism and excretion/elimination that cause toxicity. Current toxicity prediction approaches use either mechanistic or correlative methods. Correlative systems take molecular descriptors, biological data, and chemical structures and, by use of statistical analysis of data sets, represent them in mathematical models. The models describe the relationships between structure and activity and can be used to predict toxicity. The mechanistic approach involves human experts who make a considered assessment of the mechanism of interaction with a biological system, taking the molecular properties, biological data, and chemical structures into account. The correlative approach uses an unbiased assessment of the data to generate relationships and predict toxicity. It is capable of discovering potentially new SARs and, can lead to new ideas in the human assessment of mechanisms by which chemicals interact with biological systems. It is most useful for congeneric data sets or when one has a large amount of good data but little mechanistic knowledge. However, it can also generate relationships that have little chemical or biological plausibility. Results obtained are heavily dependent upon the quality of the data used to build the model. For these reasons careful validation is required for effective use of the correlative approach. The mechanistic method is based upon an understanding or hypothesis of the mechanisms of molecular interactions that determine the activity, i.e., there is some human input into the system of SAR generation. However, systems using this approach are restricted to human knowledge, being incapable of discovering new relationships automatically. As a consequence, they also have a tendency to be biased toward current ideas about mechanisms of action. The early toxicity models were based on QSAR models and were used to predict LD<sub>50</sub>, based upon various descriptors. It was also reported that QSAR models (partial least-squares (PLS), Bayesian regularized neural network) correlating IGC50 with the hydrophobicity, the logarithm of the 1 octanol/water partition coefficient, the molecular orbital properties, the lowest unoccupied molecular orbital energy (Elumo) and, maximum acceptor super delocalizability (Amax). More QSAR models are still coming forth. A representative mechanistic toxicity prediction approach was reported by Sanderson and co-workers. The program is now commercially available. Artificial neural networks (ANN) have recently been applied in toxicity predictions; these include: back-propagation neural network. Varied as these areas are and diversified as these applications are, the field of chemoinformatics is by far not fully developed. 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.

## **b. Teaching of Chemoinformatics**

Chemists have to become more efficient in planning their experiments, have to extract more knowledge from their data. Chemoinformatics can help in this endeavor. Furthermore, it is important that a certain amount of chemoinformatics is integrated into chemistry curricula in order that chemists realize where chemoinformatics could help them, where they best ask chemoinformatics experts. In addition, a few universities have to offer training for chemoinformatics specialists. The first steps have already been made at a variety of universities around the globe. More has to come in order that more experts on chemoinformatics are trained that society so urgently needs? The universities are offering courses on Chemoinformatics are University of Sheffield (Willett) - MSc/PhD programs, University of Erlangen (Gasteiger), UCSF (Kuntz), University of Texas (Pearlman), Yale (Jorgensen), University of Michigan (Crippen), Indiana University (Wiggins) - MSc program, Cambridge Unilever (Glen, Goodman, Murray-Rust), Scripps - Molecular Graphics lab, Bioinformatics Institute Of India , Chandigarh.

## **Model to predict toxicity**



Source: Clark, D. E., (2000)

## **Chemoinformatics in Textile Industry**

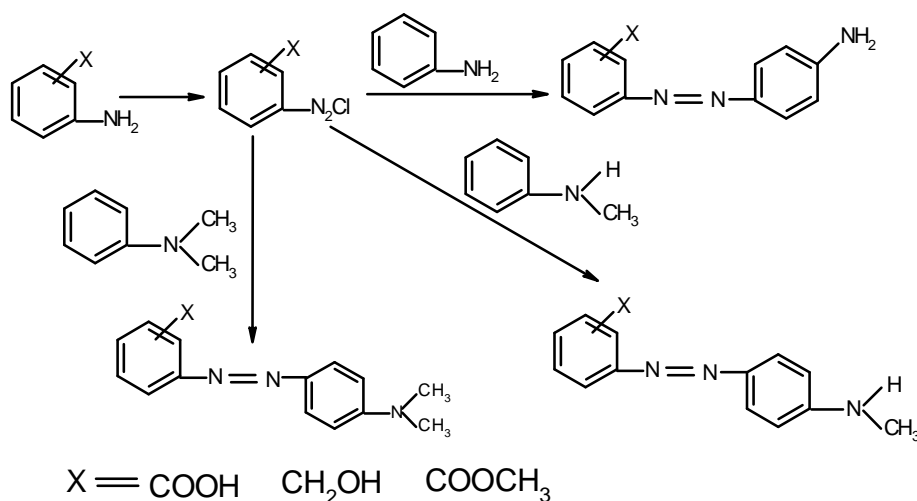
Combinatorial organic synthesis (COS), high through-put screening (HTS), and chemoinformatics (CI) are highly efficient and cost-effective tools to develop novel, state-of-the-art, non-toxic chemicals (*e.g.* dyes, colorants, finishes, pigments, surfactants, etc.) of commercial importance to the textile industry.



## Combinatorial Organic Synthesis (COS)-

Earlier a Quest 210 SLN Organic Synthesizer which can run up to twenty reactions simultaneously, integrating multi-step solution-phase synthesis, workup, initial purification and product collection. The Organic Synthesizer can heat, cool, mix, reflux, perform liquid-liquid extractions, concentrate products, etc. Now, it is integrated with a CombiFlash<sup>TM</sup> purification system which allows state-of-the-art productivity and versatility in automated organic purification systems. It is the system of choice for methods development, variable scale automation. It incorporates on-line detection, flexible sample loading and supports a PC-based method for data management; each sample can be run with customized parameters for sample size, solvent system, and gradient and flow rate.

By integrating CombiFlash<sup>TM</sup> with the organic synthesizer the combinatorial effort helps to synthesize other chemicals of importance to the textile industry. A typical library synthesized using this integrated approach is shown in Scheme 1.



### Synthesis of Azo dye

Source: Bhat *et.al.*,(2002).

## High-Throughput Screening (HTS)

HTS is the integration of technologies (laboratory automation, assay technology, micro plate based instrumentation, etc.) to quickly screen chemical compounds in search of a desired activity. While the preliminary results are promising we are keeping other options open, such as COPASTM technology which can optically analyze, sort and dispense various kinds of small diverse molecules including combinatorial beads and plant seeds.

## Chemoinformatics (CI)

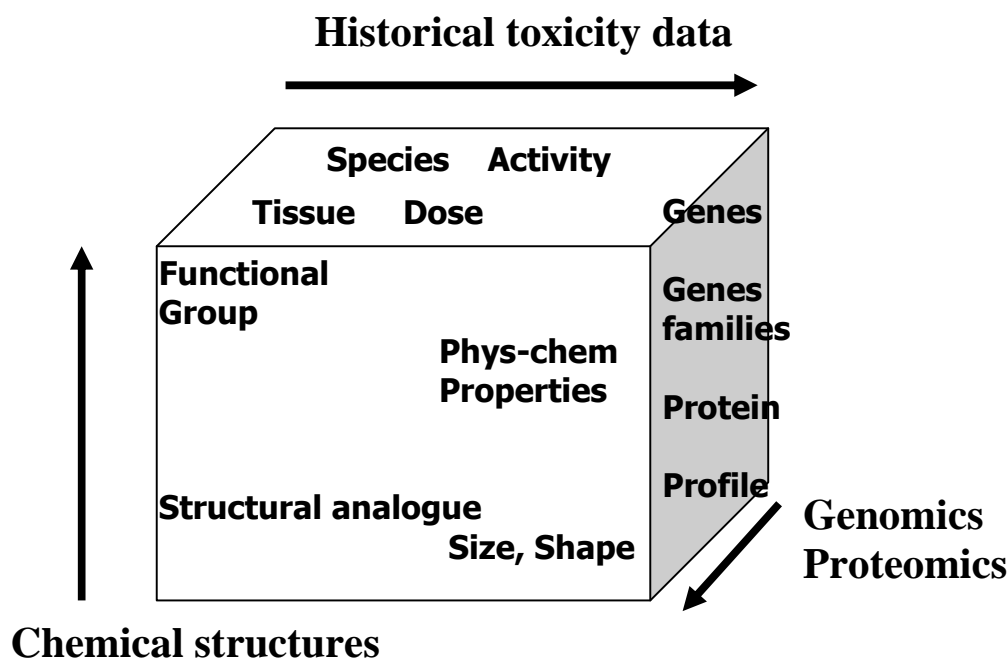
Chemoinformatics to explore differences in structural and electronic features in positional isomers is an integral part of the overall combinatorial process. Based on density functional theory calculations that there are differences in the carcinogenic behaviors of azo dyes. Quantitative structure-activity relationships that correlated the

observed mutagenic activity of 43 aminoazobenzene derivative with a variety (>300) of molecular descriptors (constitutional, topological, geometrical, electrostatic, quantum-chemical and thermodynamic) calculated using quantum-chemical semi empirical methodology.

## **Chemo bioinformatics**

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.



Source: Drews, J., (2000)

## **Key challenges**

The key challenge for computational methods then is not traveling through chemical space per se, but rather to be able to focus traveling expeditions in a vast chemical space towards interesting regions, and to be able to recognize interesting stars and galaxies

when they are encountered. The notion of what is interesting may vary of course with the task (e.g. drug discovery, reaction discovery, polymer discovery). But at the most fundamental level what is needed are tools to predict the physical, chemical, and biological properties of small molecules and reactions in order to focus searches and filter search results. Computational methods in chemistry can be organized along a spectrum ranging from Schrodinger equation, to molecular dynamics, to statistical machine learning methods. Quantum mechanical methods, or even molecular dynamics methods, are computationally intensive and do not scale well to large datasets. These methods are best applied to specific questions on focused small datasets. Statistical and machine learning methods are more likely to yield successful approaches for rapidly sifting through large datasets of chemical information. Because in the absence of large public database and datasets, chemoinformatics is in a state reminiscent of bioinformatics two or three decades ago, it may be productive to adapt the lessons learnt from bioinformatics to chemoinformatics, while maintaining also a perspective on the fundamental differences between these two relatively young interdisciplinary sciences. If this analogy is correct, two key ingredients were essential for unlocking the large-scale development of bioinformatics and the application of modern statistical machine learning methods to biological data, data and similarity measures. In bioinformatics, such as Genbank, Swissprot, and the PDB while alignment algorithms have provided robust similarity measures with their fast BLAST implementation becoming the workhorse of the field. Mutatis mutandis, the same is likely to be true in chemoinformatics.

This new drug discovery strategy, challenges cheminformatics in the following aspects:

(1) cheminformatics should be able to extract knowledge from large-scale raw HTS databases in a shorter time periods, (2) cheminformatics should be able to provide efficient in silico tools to predict ADMET properties,

## **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 cheminformatics have to be integrated into chemistry curricula, a few universities have to offer full cheminformatics 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 cheminformatics.

## **References**

- Bhat K; Bock C., Howard NJ.(2002) COS and HTS design of high-performance, non-toxic chemicals for textiles, NTC Project: C00-PH01 (formerly C00-P01)
- Brown F.K. (1998), Chemoinformatics: What is it and how does it Impact? Drug Discovery Ann. Reports Med. Chem., **33**:375-384.
- Clark, D. E. and Pickett, S. D., “Computational methods for the prediction of ‘drug likeness’”, *Drug Discov. Today*, 2000, **5**, 49-58.
- Drews J, Drug discovery: a historical perspective, Science, 287 5463: **pp**1,960-1,964, 2000
- Gasteiger J. and Funatsu K. (2006) Chemoinformatics – An Important Scientific Discipline, *J. Comput. Chem. Jpn*, **5(2)**: 53–58
- Gasteiger, Editor, Handbook of Chemoinformatics - From Data to Knowledge, Wiley-VCH, Weinheim (2003).
- Gasteiger, J. T. Engel, Editors Chemoinformatics - A Textbook, Wiley-VCH, Weinheim (2003).
- J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, 2nd Edition, Wiley-VCH, Weinheim (1999).
- Leach AR., Gillet VJ.(2003) An Introduction to Chemoinformatics, Springer:1-57
- Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. “Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings”, *Adv. Drug Deliv. Rev.*, 1997, **23**, 3-25.
- Oprea, T. I., Davis, A. M., Teague, S. J., and Leeson, P. D. “Is There a Difference between Leads and Drugs? A Historical Perspective”, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1308 -1315.
- R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, J. Lederberg, Applications of Artificial Intelligence for Organic Chemistry; the Dendral Project, McGraw-Hill, New York (1980).
- Wild J D, Getting Started in Chemoinformatics, Version 1.0, September 2004
- Woo. (1996) *Environ. Carc. & Ecotox. Rev.*, **C14**:1-42
- Xu J. and Hagler A. (2002) Chemoinformatics and Drug Discovery, *Molecules*, **7**: 566-600