ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Chemoinformatics: The Application of Informatics Methods to Solve Chemical Problems

Satish A Bhalerao*, Deepa R Verma, Rohan L D'souza, Nikhil C Teli, and Vinodkumar S Didwana

Environmental Sciences Research Laboratory, Wilson College, Mumbai-400 007
Department of Biological Sciences, VIVA College, Virar (w)-401 303

ABSTRACT

Chemoinformatics is a modern computational tool that works by bringing together both information technology and chemistry to solve problems related to drug discovery. These methods can also be used in chemical and allied industries in various other forms. Chemoinformatics gives us an opportunity to transform the data obtained via linking the two fields into knowledge which can then be extended to make proper and better decisions in areas such as drug discovery, understanding chemical interaction, standardization of drug manufacturing protocols etc. The primary application of chemoinformatics 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. Chemoinformatics can help us to carry out virtual experiments, which provides insights as to how our body might actually respond to the drug. Apart from that the technique is very fast; so that the time is saved along with it no practical working is required which is an added advantage. However everything in Chemoinformatics is hypothetical and we cannot surely say that the drug will act the way it was predicted too.As more drug discovery research is carried out in academia, institutes and small companies, and solutions will require pieces from chemoinformatics, bioinformatics and other disciplines, chemoinformatics knowledge and tools should be made as widely available as possible. All problems in chemistry require novel approaches for managing large amounts of chemical structures and data and for modelling complex relationships. This is where chemoinformatics methods can come in.

Keywords: *Chemoinformatics, drug discovery, standardization, hypothetical, bioinformatics*

***Corresponding author**

INTRODUCTION

The first computer based system was established over 40 years ago. The tem chemoinformatics was defined by F.K. Brown in 1998. With all the problems at hand in chemistry, complex relationships, profusion of data and lacks of necessary data, owing to this problem the new field of chemo informatics came into existence. Chemoinformatics is a means of bringing together chemistry and information technology for making rapid analysis without actual experimentation. One of the important applications of Chemoinformatics is the development of models linking chemical structure and various molecular properties. Thus the use of I.T and chemistry has played an important role in the area of drug discovery and organization of data. Today about 45 million chemical compounds are known and this number is increasing by several millions every year. All this data thus collected is stored in a database and can be made accessible to all. This in one way has a potential to create a revolution by making available a lot of useful information; which can be used for understanding the chemistry behind drug discovery. It is an important scientific discipline; standing on the interface between chemistry, biology and computer science [1].

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 information collected is not only useful for drug discovery, but can also be used for data analysis in industries such as paper and pulp, dye and allied industries [2]. 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 [3]. 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 [4].

This review paper deals with history, scope and fundamentals of chemoinformatics, relations of chemoinformatics with other disciplines, chemoinformatics as a tool in drug discovery and its future prospects.

MAJOR ASPECTS OF CHEMOINFORMATICS

1. **Information Acquisition and Management**: Methods for collecting data (mainly experimental). Development of databases for storage and retrieval of information.
2. **Information Use:** Data analysis, correlation and model building.
3. **Information Application:** Prediction of molecular properties relevant to drugs, virtual screening of chemical libraries, system chemical biology networks.

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 [2]. 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 [5].

RELATIONS OF CHEMOINFORMATICS WITH OTHER DISCIPLINES

Chemoinformatics and Machine learning although machine learning is widely used for structure property modeling, chemoinformatics can be considered as a very specific area of its application. The specificity of chemoinformatics results from (i) the nature of chemical objects, (ii) the complexity of the chemical universe and (iii) a possibility to take into account an extra-knowledge. The basic chemical object is a graph (or hyper graph), rather than simple fixed-sized vector of numbers as in the typical applications in mathematical statistics and machine learning. This dictates the need to apply graph theory, to develop novel descriptors and structured graph kernels, and to apply machine learning methods capable of dealing with structured discrete data.

The second important distinction comes from the fact that the chemical data result from an explorative process in a huge chemical space rather than from specially organized sampling. Hence, they cannot be considered as representative, independent and identically distributed sampling from a well defined distribution. Thus, special approaches are Chemoinformatics as a Theoretical Chemistry Discipline needed to treat this problem: various strategies to explore chemical space, the “applicability domain” concept, the active learning approach, etc. [6].

Finally, one can use the relationships between different properties issued from physicochemical theory. (For example, the Arrhenius law could be particularly useful upon the modeling the rate constants). These relationships could be integrated into chemoinformatics workflow as an external knowledge.

1. Chemoinformatics and Chemmetrics

Massart [7] has defined chemometrics as "a chemical discipline that applies mathematics, statistics and formal logic (a) to design and select optimal experimental procedures; (b) to provide maximum relevant chemical information by analyzing chemical data; and (c) to obtain knowledge about chemical systems". Generally, chemometrics requires no information about chemical structure and, therefore it overlaps with chemoinformatics only in the area of application of machine learning methods. It is widely used in experiment design, chemical engineering, analytical chemistry and treatment of spectra – fields where an exhaustive treatment of multivariate data is needed.

2. Chemoinformatics and Bioinformatics

Unlike chemoinformatics dealing with "chemical size" molecules, bioinformatics uses computational tools to study the structure and function of biomolecules (proteins, nucleic acids). This is a broad field mostly involving 3D (force field and quantum mechanics calculations) and 1D (sequence alignment) modeling. In the latter, a biomolecule is represented as a string of characters (building blocks). Graph and fixed size vector models used in chemoinformatics are very rarely used in bioinformatics. In this sense, chemo and bioinformatics are "complementary". On the other hand, there are many examples of interpenetration of these fields. Thus, in docking calculations, protein structures could be generated by bioinformatics tools, whereas some scoring functions involve vector representation of ligands. Another way to combine bio- and cheminformatic approaches is related to the construction of protein-ligand descriptors or fingerprints based on available 3D information about protein-ligand complexes. Thus, Tropsha et al. developed CoLiBRI descriptors calculated for a pseudo molecule constructed from interacting atoms of the protein and the ligand. Marcou and Rognan [8] have developed "interaction fingerprints" accounting for eight interaction types per each protein atom interacting with the ligand: hydrophobic; aromatic (face to face); aromatic (edge to face); H-bond (protein donor atom); H-bond (protein acceptor atom); ionic (positively charged protein atom); ionic (negatively charged protein atom); metal complexation., Langer et al.[9] have reported a technique to build pharmacophoric ligand models based on the analysis of 3D protein-ligand structures.

A promising way to describe ligand–receptor complexes concerns construction of protein-ligand kernels (PLK) as products of "chemical" ligand–ligand (LLK) and "biological" protein–protein kernels (PPK). The resulting feature space for PLK is a tensor product of the features spaces corresponding to LLK and PPK. Machine learning models involving PLK are based on the idea that similar ligands bind to similar proteins. Using these kernels, one can predict binding potency of both different ligands with respect to a given protein, and different proteins with respect to a given ligand. Several articles describing PPK have been published. Erhan et al. combined "chemical" kernels based on MOE descriptors and "biological" kernels based on protein- ligand "interaction fingerprints". Faulon et al. used the signature molecular descriptors to calculate "chemical" and "biological" Tanimoto kernels. Jacob and Vert [10] combined a Tanimoto kernel for the ligands and several types of kernels for the proteins. In particular, for PPK they compared either protein sequences or EC numbers. Bajorath et al. used a linear kernel for the ligands and protein-protein kernels calculated from sequence identity matrix.

APPLICATIONS OF CHEMOINFORMATICS

I) Fields of Chemistry

The range of applications of chemoinformatics is rich indeed; many 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

a) Storage and retrieval of chemical structures and associated data to manage the flood Data
b) Dissemination of data on the internet
c) 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

a) Elucidation of the structure of a compound based on spectroscopic data

4. Organic Chemistry

a) Prediction of the course and products of organic reactions
b) Design of organic syntheses

5. Drug Design

a) Identification of new lead structures
b) Optimization of lead structures
c) Establishment of quantitative structure activity relationships
d) Comparison of chemical libraries
e) Definition and analysis of structural diversity planning of chemical libraries
f) Analysis of high-throughput data
g) Docking of a ligand into a receptor
h) *de novo* design of ligands

i) Modeling of ADME-Tox properties
j) Prediction of the metabolism of xenobiotics
k) Analysis of biochemical pathways 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.

II) Teaching 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 needed?

NEED OF CHEMINFORMATICS 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 [8].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 [11].

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 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.

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.

3. Data in Chemistry

Much of our chemical knowledge has been derived from data. Chemistry offer a rich range of data on physical, chemical, and biological properties: binary data for http://www.sccj.net/publications/JCCJ/ 55 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. Data sources 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.

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.

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 additivity 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.

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.

FUNDAMENTALS OF CHEMOINFORMATICS

For the objects in chemical space, chemoinformatics builds its models using two main mathematical approaches: graph theory and statistical learning. While these mathematical methods can be applied to other fields, the chemical space is a particular concept of chemoinformatics describing a way to handle ensembles of chemical structures.

1. 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’s for molecular modeling are ArgusLab, Chimera and Chemical.

2. 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 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.

3. 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.

4. Chemical Space Paradigm

As pointed out by C. Lipinski and A. Hopkins, “chemical space can be viewed as being analogous to the cosmological universe in its vastness, with chemical compounds populating space instead of stars” [4]. Any attempt even to count the number of chemical compounds which potentially could be synthesized leads to combinatorial explosion and yields an absolutely unrealistic number estimated as more than 1060 [2] which exceeds the number of elemental particles in the cosmological universe. Clearly that this number is so huge that it is impossible not only to synthesize these molecules but even to generate computationally their structures. The goal of chemoinformatics is to find a rational way of representing this literally infinite chemical space and to navigate in this space. Efficient strategies for navigating chemical space are crucially important for the development of new biologically active compounds and the design of new drugs for medicine [4]. This is due to the fact that biologically active compounds of a certain type are not distributed evenly over the whole chemical space, but form very compact regions in it, like galaxies in the cosmological universe [4].This is certainly true for any other chemical property. A special term, chemography, analogous to geography, has even been suggested for the art of navigating in chemical space [12]. Although the expression “Chemical space” is widely used in the chemoinformatics literature, it is not still well defined. Generally speaking, the notion of “space” stands for a set of objects with some particular properties and some relationships between them (metric).

a) Representation of Chemical Objects in Chemoinformatics

In chemoinformatics, the molecules are treated as informational objects, identifying their structure and properties. Generally, two main types of objects are used: graphs and descriptor vectors. In a vertex- and edge-labeled undirected graph, the vertices and edges correspond to atoms and chemical bonds, respectively. The vertex labels identify symbols of chemical elements, whereas the edge labels characterize the bond type. The label corresponds either to the bond order in molecules or to some special bond types in more complex systems. For instance, different types of “coordination” bonds can be defined for supramolecular systems, whereas “dynamic” bonds corresponding to chemical transformations can be used to encode chemical reactions [11]. More complex chemical systems, like polymers or mixtures can be described by ensembles of graphs. For several practical purposes, more generalized representations of chemical structures are needed. For example, for pharmacophore analysis, the graph vertexes can be labeled as pharmacophoric centers (H-donors, H-acceptors, cation, anion, aliphatic, aromatic), while the separation of two centers can be depicted by an edge labeled by the value of the 2D or 3D distance [8]. In Markush structures used for patent searches, a graph vertex can stand for several types of either individual atoms or whole substructures (e.g., substituents). The same is true for substructure queries used for searching chemical databases [13]. Consideration of some complex chemical objects reveals, however, some limitations of graph theory to code chemical structures and their ensembles. Instead, hyper graphs [14] have been suggested as a more adequate mathematical model to encode stereo chemical information and multicenter bonds. However, hyper graphs are much more difficult objects to operate compared to graphs, and, therefore, their use is still very limited. Another popular representation of molecular structure is based on molecular descriptors defined by to deschini and Consonni as "…the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment." [15] This molecular representation is extremely popular in chemoinformatics because: (a) various descriptors can be generated from one and the same molecular graph, Thus describing different facets of the information hidden in the graph; (b) it is invariant to any renumbering of graph vertices; (c) most of the descriptors are easy interpretable; (d) inductive transfer of knowledge can be performed via descriptors;[11] and, (e) descriptors define a vector space which is mathematically much easier to handle compared to the graph-based space. Descriptor vectors can be prepared not only for individual molecules but for more complex systems like chemical reactions [11] or multi component mixtures [16]. Nowadays, more than 5000 types of descriptors of different types have been reported [15]. They are used for database processing (as screens or fingerprints), for building SAR/QSAR/QSPR models, in similarity searching, clustering, etc. At the same time, several weak points of molecular descriptors should be mentioned: (a) If descriptors are not well selected, in the resulting chemical space two different molecules can be superposed on one point; (b) The number of existing descriptors is very large and despite numerous variables selection techniques reported in the literature,[17] there is always a risk of selecting irrelevant and redundant descriptors; (c) A serious drawback of molecular descriptors is the loss of reciprocity with the molecular structure. Indeed, the reverse reconstruction of molecular graphs from descriptors is a very difficult and, in some cases, impossible task known in QSAR as the "inverse" problem. From the practical point of view, it concerns generation of molecular structures possessing desired property values. Attempts to solve this problem have been reported by Gordeeva et al., Skvortsova et al. and Faulon et al. who observed some degeneracy of solutions, when several chemical structures corresponded to one set of molecular descriptor values. As pointed out in [18], this prevents a reverse engineering of chemical structures from molecular descriptors, but, on the other hand, can be useful to safely exchange chemical information in the form of molecular descriptors.

b) Chemical Similarity as a Metric of Chemical Space

By definition, a metric is a function which defines a distance between the elements of a set. For all x, y, z, this function must satisfy the following conditions: (i) d(x, y) _0 (no negativity); (ii) d(x, y) =d(y, x) (symmetry) and, (iii) d(x, z) _d(x, y) +d(y, z) (triangle inequality). Strictly speaking, the distance d(x, z) is a dissimilarity measure which is zero for identical elements and increases with the decrease of similarity between them. Thus, it can be defined as distance= 1 similarity. Some similarity measures are briefly considered below. Molecular similarity (or chemical similarity) is one of the most basic concepts in chemoinformatics [19]. It is widely used in virtual screening and in silico design of new compounds. Such studies are based on the similar property principle which states that similar compounds have similar properties [19]. In application to classification problems this means that similar chemical compounds tend to belong to the same class (e.g., possessing similar biological activity), whereas as applied to regression problems it means that the approximating function should be as smooth as possible. It should also be pointed out that molecular similarity always depends on the choice of descriptors and methods to compare molecular graphs. Chemical similarity measures described in the literature can be calculated from (a) molecular graphs; (b) descriptor vectors; (c) molecular fields; they can also be assessed from (d) kernels, and (e) unsupervised or (f) supervised modeling studies. This classification is rather fuzzy, and some similarity measures belong simultaneously to several classes.

A similarity measure based on the size of the maximum common sub graph (MCS) for a pair of graphs is perhaps the most well-known graph-based similarity measure. Due to the relative complexity and inefficiency of computational algorithms to search for an MCS [20], this approach, however, is rarely used to perform a similarity search [21] or to cluster chemical databases. Another type of graph-based similarity measure is that of graph kernels which assign to each pair of graphs a positive real number characterizing similarity. They are used to map a graph-based chemical space to a vector (feature) space in which the structure–property model is built. This approach has been successfully used in SAR and QSAR [22]. The most popular similarity measures are based on fixed sized descriptor vectors. These are various types of distances (Euclidean, Manhattan, Mahalanobis, Minkowski) measuring molecular dissimilarity or some indices (Tanimoto, Dice, cosine, Tversky, etc.) measuring similarity. These measures are widely discussed in the literature, e.g., see the review paper by Willett [23] and references therein.

Several approaches have been developed to compare molecular fields. The Carbo index is computed by integrating overlaps of electronic densities of two molecules assessed using quantum-chemical approaches [24]. The SEAL index [25] is used to assess an alignment of steric and electrostatic fields of the molecules. Since any molecular field could be represented as a descriptor vector based on the field value on the grid points, a similarity measure can be simply calculated as the product of two vectors. Similarity measures for which all matrices of values are semi positive definite (the determinant is larger or equal to zero) are called “Mercer kernels”, or simply “kernels”. Generally, kernels are used to project the objects (graphs or vectors) into a Hilbert “feature space”, in which a similarity measure between these objects is equal to dot-product of their projections. A dot product of vectors, which can be viewed as the cosine similarity measure for normalized vectors, is the simplest type of kernel.

Unsupervised machine-learning methods of nonlinear neighborhood-preserving projections of data can also be used to assess similarity. A typical example is mapping to Self-Organizing (Kohonen) Maps, SOM [26], where the similarity is measured as a distance between different cells. This offers the possibility to use SOMs for property predictions [27] and in virtual screening [28]. If several QSAR models are simultaneously applied to predict a property for a series of compounds, the similarity can be assessed in the “models’ space”. Indeed, for each compound, one can form a vector based on the prediction results. A dot product of these vectors can be considered as a measure of the similarity of two molecules. This approach has been used by Tetko in the ASNN (Associative Neural Networks) method [29]. Generally, similarity measures could be used both for similarity-based predictions and similarity searching [19]. Similarity- based prediction approaches in the initial descriptor space are based on the k nearest neighbor’s method (kNN). However, kernel similarity measures implemented in kernel based machine learning methods lead generally to more Chemoinformatics as Theoretical Chemistry Discipline computationally efficient and predictive models. Both in similarity-based prediction methods and in querying large chemical databases, the computational efficiency largely depends on whether a given similarity measure defines a metric in chemical space [30].For most of similarity measures, the metric axioms (i)–(iii) are valid, and, therefore, they can be perceived as distances in chemical space.

c) Navigation in Graph-based Chemical Space

In principle, each ensemble of molecular graphs forms a discrete metric topological space. Its topology is defined by a set of all its possible subsets, where the simplest discrete metric gives the distance 0 if two chemical objects are equivalent, i.e. corresponding chemical graphs are isomorphic to each other and 1 otherwise. This simplest metric is however not useful in practical applications, because in such space all distinct objects are equally similar to each other. More flexible relationship between graphs can be expressed as a degree of their mutual similarity/dissimilarity. In particular, this relationship can be established by mapping an ensemble of graphs onto a descriptor vector space followed by an assessment of standard similarity measures.

The three main approaches used to describe a set of molecular graphs and to navigate in this space are: (a) substructure- based, (b) superstructure-based, and (c) mutation based. In the substructure-based approach a special “navigation” graph is usually constructed. It can be used for the visualization of chemical databases, exploring relations between compounds and discovering unexplored regions in the chemical space. In the navigation graph, the nodes correspond to individual molecular graphs and edges correspond to some transition rules. Bemis and Mursko have considered transitions between an unlabelled graph (framework) to a labeled graph (full chemical structure) [30]. They invented the concept of molecular frameworks used to organize the structural data by grouping the atoms of each drug molecule into ring, linker, framework, and side chain atoms. Thus, a huge database can be described by a limited number of frameworks. In the “scaffold tree” graph approach of Schuffenhauer et al. [31], transitions are allowed between a molecular graph and its sub graph. It has been demonstrated that this type of navigation graphs allows one to perform an efficient and intuitive activity mapping, visualization and navigation of the chemical space defined by a given library, which in turn leads to building correlations with bioactivity and further compound design [32]. Thus, the hierarchical scaffold classification proposed in [33] helps to chart biologically relevant chemical space using data on natural products. The idea of “scaffold tree “is implemented in the open source “Scaffold Hunter” software [34], an interactive tool for navigation in chemical space, which facilitates recognition of complex structural relationships associated with bioactivity. To represent relationships in analogous series of compounds having the same scaffold and different substitution patterns, multilayer-rooted “combinatorial analogue graphs” (CAGs) have been proposed by Peltason et al.

These graphical representations hierarchically organize compounds according to substitution patterns and are annotated with SARI discontinuity scores [35] in order to account for SAR discontinuity at the level of functional groups. The approach makes it possible to identify under sampled regions and highlight key substitution patterns which determine the SAR of a compound series. An alternative way to visualize SARs in analogous series with a common scaffold is offered by the “SAR maps” invented by Agrafiotis et al. In a “SAR map”, each series is rendered as a rectangular matrix of cells, each representing a unique combination of substituents (i.e., a unique compound). Color-coding the cells by their potency easily identifies SAR patterns. Pollock et al. introduced the scaffold topology approach, which represents a connected graph with the minimum number of nodes and edges required to fully describe its ring structure. An algorithm for systematic generation of scaffold topologies allows one to analyze systematically all scaffold topologies for up to eight-ring molecules and four-valence atoms, thus providing coverage of the lower portion of the chemical space of small molecules [36].

Scaffold topology distributions were analyzed for several of the most popular chemical structure databases with huge number of compounds, both real and virtual, and many interesting features were found [37]. It is claimed that “scaffold topologies can be the first step toward an efficient coarse-grained classification scheme of the molecules found in chemical databases”. In the superstructure-based approach, each individual molecular graph is considered as a sub graph of a common super graph corresponding to the ensemble of individual graphs [38]. Although this approach is limited to relatively small congeneric sets of compounds, it has been found very suitable to build QSAR models, as demonstrated in the positional analysis by Magee, the DARC/CALPHI system by Mercier et al. , the MTD-PLS approach of Kurunczi et al. , and the MFTA approach by Palyulin et al. For each individual chemical structure, the occupancies of super graph nodes or local physicochemical descriptors of atoms matching these nodes, form a fixed-size descriptor vector used in machine-learning methods as an input. An alternative mutation-based approach to travel in graph-based chemical space has been suggested by van Deursen et al. They represent a chemical space as a graph in which vertices correspond to individual molecules and edges correspond to structural mutations: change of atom type; inversion of stereo chemical configuration at chiral centers, removal and addition of atom; saturation and unsaturation of bond; bond rearrangement; and aromatic ring addition. Traveling in such space from one active molecule to another one, one can discover along the trajectory a certain number of novel structures which can be further analyzed in the context of lead optimization. A similar approach has been reported by Bishop et al. who suggested the use of chemical reactions as structural mutations connecting in the chemical space known organic compounds taken from the Beilstein database. The super graph created in such a way enabled the authors to select a set of the "most useful compounds" from which the majority of chemical compounds can be synthesized.

d) Navigation in Descriptor-Based Chemical Space

Descriptor-based chemical space is a multidimensional space in which molecules are represented as vectors. Two main approaches – dimensionality reduction and clustering -are used to facilitate the navigation in this space.

Dimensionality reduction is achieved in classical multivariate data analysis by the Principal Component Analysis (PCA) procedure [39]. In PCA, several features (called "principal components") corresponding to the principal inertia axes of the "cloud" of data points in the initial descriptor space are used as axes of a new low-dimensional space, onto which the initial data points are projected. Such projection occurs with the minimal loss of information and, therefore, maximal conservation of the neighborhood relationships between data points. Thus, representation of the data points in the resulting low-dimensional space can be considered as a "navigation map" of the descriptor space. This idea has been implemented in the ChemGPS (chemical global positioning system) technique which positions chemical structures in drug-like chemical space (drug space). This makes this approach as well as the related ChemGPS-NP [40] tool a well-suited reference system to compare multiple libraries and to keep track of previously explored regions of the chemical descriptor space. Although the axes of the PCA "navigation map" are orthogonal, corresponding latent variables are statistically independent only for a Gaussian distribution of data points. Since this distribution in the descriptor space is usually strongly non-Gaussian, this can hamper the chemical interpretability of particular latent variables and reduce the usefulness of the whole "navigation map". To solve this problem, Independent Component Analysis (ICA) has been suggested. It has been demonstrated that the application of ICA instead of PCA yields chemically more readily interpretable latent variables [41]. Hierarchical cluster analysis represents an alternative approach to navigate in the descriptor space. The resulting dendrogram gives a clear picture of the neighborhood relations between chemical objects, although for a large number of compounds it becomes too burdensome. The combined application of dimensionality reduction and clustering methods is realized in Kohonen Self-Organizing Maps (SOM). In SOMs, the dimensionality reduction is achieved by embedding a net of neurons onto a 2D surface. The SOMs provide more efficient solutions than PCA, because the former are more suitable to analyze complex topological structures of the descriptor space. The ability of SOMs to build "navigation maps" for visualizing chemical space has been demonstrated on GPCR ligands [34], toxic compounds [42], inhibitors of P-glycoprotein and different organic reactions [43]. A set of chemical structures can be presented as a graph in which the vertices correspond to individual molecules and the edges connecting them correspond to certain neighborhood relations. This technique has been used to represent relationships between different classes of drug molecules, to elucidate similarity relationships within the sets of active compounds [34], and to explore structure-selectivity relationships [35].

Hierarchical clustering techniques using some similarity measures also offer the possibility of analyzing large chemical data sets. Thus, Agrafiotis et al. have used radial clusterograms, different segments of which are color-coded by biological activity or any other user-defined property. To characterize structure–activity landscapes in the descriptor- based chemical space, SARI and SALI indices have been suggested. The SARI index globally characterizes structure-activity landscapes. It consists of two terms: the continuity score which measures the potency-weighted structural diversity, and the discontinuity score calculated as the average potency difference among similar pairs of molecules. The SALI index [20] is local, considering two related molecules, and it is often used to quantify “activity cliffs” [44].

5. Modeling Background

The two main mathematical approaches used in chemoinformatics are graph theory and computational learning theory. Whilst the chemical applications of graphs are described in numerous books and review articles, the latter is described mostly in the data mining literature. Here, we give some general information about some basic concepts of computational learning theory.

a) Computational Learning Theory

In recent years, in statistical modeling there has been a shift from the classical statistical paradigm of “model parameterization” to a new paradigm of “predictive flexible modeling”. The first paradigm supposes that the functional dependence between the input and output data is established from some external knowledge and the goal of the statistical study is to find a few independent free parameters by fitting to experimental data. This usually requires a certain number of experimental observations per each free parameter. Unfortunately, this requirement can be met only in very few cases, e.g., within the classical Hansch-Fujita approach based on three descriptors only [45]. The aim of the second paradigm is to build models with maximal predictive performance by fitting to experimental data rather flexible families of functions involving large numbers of intercorrelated parameters. Such a setup is evidently much more appropriate for most chemoinformatics studies. The first attempts to implement the second paradigm in the framework of so-called nonparametric statistical analysis failed because of the “curse of dimensionality” (which required a huge number of observations exponentially growing with the number of free parameters) [46]. Nonetheless, early works on predictive modeling were successfully carried out using completely heuristic methodologies of artificial neural networks [47] and decision trees [48]. For the first time, a strong theoretical background to build statistical models using finite (even small) data sets was developed by Vapnik in his Statistical Learning Theory (SLT). This approach, together with that developed later as the PAC (Probably Approximately Correct) theory by Valiant and the MDL (Minimum Description Length) concept by Rissanen constitute the basis of modern computational learning theory. According to SLT, the goal of statistical study is to choose from a given set of functions f(x, q) the “best” one f(x, q*) with the minimum value of the risk functional R[f], which is defined as an expected prediction error on new data taken from the same distribution as the training set (i.e., the mean prediction performance on all possible test sets). Here x denotes the variables (descriptors in QSAR studies) and q the adjustable parameters. Another important characteristic is the empirical risk functional Remp [f], which is defined as an error on the training set (fitting error). For regression tasks, one of the most interesting conclusions of SLT is that the value of the complexity term does not directly depend. On the number of free parameters q in the function class f, the flexibility (capacity, complexity) of which is measured by the VC dimension h. The value of h can be considered as an “effective” number of free parameters. (Note that h is equal to the number of free parameters in classical multiple linear regression without descriptor selection).

According to SLT, h is controlled by the trade-off parameter used to simultaneously minimize both terms in Equation 2. This offers an opportunity to build models with any (even very huge) number of variables using kernel approaches, which approximate nonlinear functional dependencies of any form by projecting descriptors onto a feature space of any (even infinite) dimensionality and build linear models in this feature space.

Nowadays, computation learning theory represents a quickly developing area. Thus recently, a Bayesian learning approach to predictive flexible modeling has been described [49] Instead of one single model (as in STL), it considers the whole statistical distributions of models weighted by their ability to fit data, thus allowing one to make probabilistic predictions by averaging these distributions. This approach has come to be rather popular in chemoinformatics: its implementations in Bayesian Neural Networks, [50] Gaussian Processes [51] and Bayesian Networks [52] have been recently published.

b) Different Facets of Statistical Modeling

It should be pointed out that the range of application of different statistical (machine learning) methods in chemoinformatics is currently very wide. Most of the existing machine learning approaches can provisionally be divided into two large families: supervised and unsupervised machine learning. (Some other approaches – semi supervised, active and multi-instant learning – are very rarely used in chemistry so far).

The goal of the supervised learning in chemistry is to predict physicochemical properties and biological activities of chemical compounds. The quantitative prediction of real-valued properties is performed by regression models, whereas qualitative predictions (“active” or “inactive”?) are assessed in classification models. The most popular regression methods currently used in chemoinformatics applications are multiple linear regression (MLR), partial least squares (PLS), neural networks, support vector regression(SVR), and kNN, whereas the na_ve Bayes, support vector machines (SVM), neural networks and classification trees (especially the Random Forest method[53] are widely used for classification. There are also ranking models [54] in which ranking order instead of property values are predicted, and models with structured output [47] in which predicted values belong to classes of any complexity. Models of the latter two types can be built using some special modifications of SVM. Unsupervised learning describes the data and reveals their hidden patterns. The most important tasks treated by unsupervised modeling approaches are: (a) cluster analysis (data reduction); (b) dimensionality reduction; (c) novelty (outlier) detection. All these tasks can be perceived as particular cases of data density estimation. Many standard algorithms for both nonhierarchical (e.g., k-means) and hierarchical clustering algorithms are used. The most popular algorithms for dimensionality reduction are PCA (Principal Component Analysis) and ICA (Independent Component Analysis). Tasks (a) and (b) are solved simultaneously in the Kohonen Self-Organizing Maps (SOMs) [47], which are intensively used for the purposes of visualization and analysis of the chemical space. The ability of several machine learning methods, such as one-class SVM, to tackle the problem of novelty detection is currently used to define the applicability domains of QSAR/QSPR models [55] as well as in virtual screening experiments [47]. With respect to data description, two types of models – primal and dual – can be identified. Primal models are based on the direct use of descriptors, whereas dual models are based on measures describing similarity relationships between chemical structures. Kernels represent the most useful types of such measures; they can be computed both from molecular descriptors and by direct comparison of chemical structures. Both primal and dual approaches can be used within supervised and unsupervised modeling tasks. Finally, statistical models can be built for a net of mutually related models, in which their predictive performance can be leveraged due to Inductive Learning Transfer phenomenon [11], in the framework of the Multi-Task Learning and Feature Net approaches.

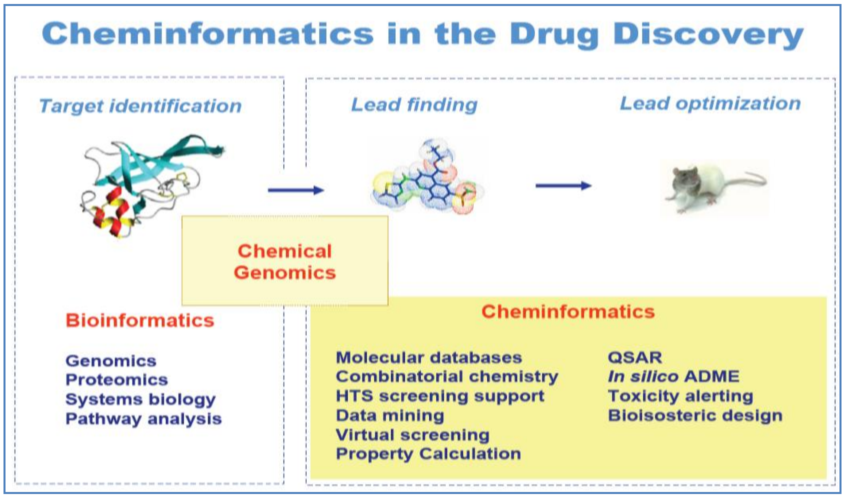
CHEMOINFORMATICS TOOL IN DRUG DISCOVERY

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. 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 [46].

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.



Source: Cheminformatics in Modern Drug Discovery Process, Peter Ertl

Figure 1: Steps towards drug discovery using chemoinformatics as a major tool.

1. **Data:**

Large and well-annotated datasets are essential for developing statistical machine learning methods in chemoinformatics, whether supervised or unsupervised, including predictive classification, regression, and clustering of small molecules and their properties.

Several parallel support have emerged recently to start to address the data bottleneck, including PubChem (http://pubchem.ncbi.nlm.nih.gov), the Harvard Chembank, UCSF's ZINC, and the UCI ChemDB [56]. The UCI ChemDB is a public database containing over 4M compounds as well as a repository of annotated datasets that can be used to develop statistical machine learning methods. Together, these datasets already pose important challenges for both supervised and unsupervised machine learning methods, from clustering to kernel methods.

2. **Drug Discovery:**

There are seven steps in the drug discovery process: disease selection, target hypothesis, lead compound identification (screening), lead optimization, pre-clinical trial, and clinical trial and pharmacogenomic optimization. Traditionally, these steps are carried out sequentially. The average cost of creating a NCE in a major pharmaceutical company was estimated at around $7,500/compound. To reduce costs, pharmaceutical companies have had to find new technologies to replace the old “hand-crafted” synthesis and testing NCE approaches. Since 1980, with the advent of high throughput screening (HTS), automated techniques have made possible robotized screening. Through this process, hundreds of thousands of individual compounds can be screened per drug target per year. In response to the increased demand for new compounds by biologists, chemists started using combinatorial chemical technologies to produce more new compounds in shorter periods. Combinatorial chemistry (CC) systematically and repetitively yields a large array of compounds from sets of different types of reagents, called "building blocks". By 2000, many solution- and solid-phase CC strategies were well-developed [57].

3. Drug and Lead properties:

Technologies have been developed to recognize drug-like compounds from a diverse compound library [58]. These drug-like measuring and filtering technologies have partly solved the screening problems. However, they have not been good enough to completely solve these problems. It has been observed that many drug-like compounds, which should be potential candidates; do not come up as hits when they are screened against biological targets. It is believed that further refinement of the filtering technologies should be made in order to recognize *lead-like* compounds [27] instead of *drug-like* compounds. Intrinsically, lead-likeness and drug likeness are the descriptors of potency, selectivity, absorption, distribution, metabolism, toxicity, and scalability

4. Drug Discovery Process and Early ADMET Prediction:

One now finds too many hits when searching for lead candidates, thus lead optimization is stymied. To get more target structural information, high throughput protein crystallization has been explored. Lead optimization remains the most serious bottleneck. In addition, we know that, about forty percent of all development candidates fail due to absorption, distribution, metabolism, excretion and toxicity ("ADMET") problems [30]. HTS for pharmaceutical discovery was used as a filter in order to identify the few potentially promising hits in a corporation's synthetic archive. Therefore, HTS data analyses were focused on hits, and the bulk of the non-hit data was ignored [30]. Cheminformatics methods must be applied while generating data using high throughput techniques in order to assure that good ADMET properties are achieved while making and screening compounds, this approach is called a multi-parametric optimization strategy [59].

5. Chemical structure database:

Since structure and substructure searches are typical NP problems, they were computationally costly [31]. In order to make structure and sub-structure searching feasible on slow computer systems, many methods were attempted in order to find concise structural representations, such as, linear notations. These convert structural graphs to strings that can easily be searched by a computer. The data screening strategies filtered out the compounds were not the main structural features (search keys) in a given query. Then, an atom-by-atom search algorithm was applied (this was usually time consuming) to a smaller number of compounds. Subsequently, screening approaches have been used in most of chemical database management systems.

6. 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 nonexperts to understand. Later, David Weininger suggested a new linear notation designated as SMILESTM [32]. SMILESTM is widely accepted and used in many chemical database systems.

7. Visualizing structures from graphed data points:

Chemical structure graphs are chemists' natural language. Since a compound library is mapped to points on a two dimensional graph, a reasonable requirement is for one to have an easy way to see the structure by pointing to the corresponding dot. This problem has been well resolved by Spotfire® software. The criteria used for selecting descriptors should be: (1) the selected descriptors should be bioactivity related (requiring correlation analysis), (2) the selected descriptors should be informative (should have diversified value distributions), (3) the selected descriptors should be independent of each other (if two descriptors are correlated to each other, related property will be unfairly biased), (4) the selected descriptors should be simple to extract, easy to explain to a chemist, invariant to irrelevant transformations, insensitive to noise, and efficient to discriminate patterns in different categories (specificity). After comparing performance and predictability in high throughput data mining, researchers from multiple groups have consistently concluded that 2D descriptors perform significantly better than 3D descriptors [38].

8. Clustering and Partitioning:

The term cluster analysis (CA) was first used by Tryon, in 1939. Actually CA encompasses a number of different classification algorithms. A general question in many areas of an inquiry is how to organize the observed data into meaningful structures, that is, how to develop taxonomies [60]. Hierarchical clustering rearranges objects in a tree-structure. Javis-Patrick (also known as nearest neighbor cluster algorithm) is commonly used to cluster chemical structures [61]. One of the most popular decision tree techniques is recursive partitioning (RP). It has been reported that RP algorithms can partition on data sets with over 100,000 compounds and 2,000,000 descriptors, in less than an hour. RP algorithms can also be used to build multivariable regression models.

9. Virtual library generation and Virtual screening:

As equipment is being automated and miniaturized, HTS capacity keeps expanding. But, increased HTS efforts have not significantly increased drug discovery successes. Considering total lead-like molecular space, the total percentage of compounds that current technologies have made and screened, is still small. This has made way for the birth of *in silico* or virtual screening (VS) technology [39]. In conjunction with high-throughput screening technology, virtual screening has become a main tool for identifying leads [62]. Virtual screening is actually one of the computational tools used to filter out unwanted compounds from physical libraries or *in silico* libraries. In order to reduce drug discovery costs, one needs to remove undesired compounds as early as possible. Filters have been built based upon oral bioavailability, aqueous solubility, and metabolic clearance and, chemically reactivity or toxic chemical groups [40]. A virtual screening method for identification of "frequent hitters" in compound libraries has been reported. If the target structure is known, one of the structure-based virtual screening methods that can be used is high throughput docking. If the target structure is unknown, but the ligands from the literature or, competitors are known, then, similarity approaches can be applied. If neither target structure nor ligand structure is known, then SAR patterns can be derived from experimental screening data by statistical approaches [63] Also, virtual screening is a great tool for the design of a combinatorial library with a given target. For example, Hopfinger and co works have constructed a combinatorial library of glucose inhibitors of glycogen phosphorylase *b* using virtual screening technology and 4D-QSAR analyses [64]. Using the 4D-QSAR model developed for a training set of 47 glucose analogue inhibitors of glycogen phosphorylase, the investigators have developed a virtual approach to screen a focused combinatorial virtual library of 225 inhibitors. Analysis of the binding predictions across the virtual library reveals patterns of structure activity information. The patterns are then used to design new focused libraries. A recent review has indicated that HTS and VS are moving toward integration [65].

10. In silico ADMET:

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 [34]. The method used (VolSurf) transforms 3D fields into descriptors and correlates them to the experimental permeation by a discriminate partial least squares procedure [35]. 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. Binding to HSA turns out to be determined by a combination of hydrophobic forces together with some modulating shape factors [20]. This agrees with X-ray structures of HAS alone or, bound to ligands, where the binding pockets of both sites and II are composed mainly of hydrophobic residues [44]. HTS has been used for metabolism and pharmacokinetics [45]. *In vitro* approaches determine metabolic stability, screening for inhibitors of specific cytochrome P450 isozymes and, identifying the most important metabolites, QSAR and pharmacophore models, protein models, and expert systems. QSAR and pharmacophore models predict substrates and inhibitors of a specific cytochrome P450 isozyme [47]. 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. 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 [51].

CHEMOINFORMATICS –FUTURE TRENDS

1. Global databases, integration of multiple data sources, public (Wikipedia-like) curation
2. 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.
3. Chemoinformatics methods will be extended to theoretical chemistry, stimulation of reaction; study of proteins will be the future areas of thrust for chemoinformatics.
4. Use of large chemo genomics databases (WOMBAT, GVK …)
5. Text and image mining, automatic extraction of useful information from publications and patents
6. Integration with bioinformatics, with focus on ligand protein interactions and pharmacophores
7. Disappearing border between cheminformatics and computational chemistry
8. In technology area –modularization, web services
9. Open source collaborative software development
10. Using all the advanced chemoinformatics system, it enhances the drug discovery rapidly and with low cost and helps to eminent scientists to synthesize the chemical molecules which lead to helps the society.

CONCLUSION

Chemoinformatics can hence be described as the application of informatics methods to solve chemical problems. It has developed over the last 40 years to a mature discipline that has applications in many areas of chemistry. It is an important scientific discipline that stands on the interface between chemistry, biology and Information Technology. Chemoinformatics spans a very broad range of problems and approaches which are often inter-related and sometimes difficult to categorize. 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, Chemoinformatics will further introduce advancements in chemistry 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. We believe that this review will be the defining theme and might help to provide much new advancement in the field of chemoinformatics in coming years. Hopefully, the availability of information related to chemoinformatics will catalyze further advancements and would open new advancements in this field.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to: Dr. V.J. Sirwaiya, Principal, Wilson College; Management, VIVA Trust, Virar (W) Maharashtra. K.P.N. Kutty, Co-Ordinator, VIVA Trust, Virar (W) Maharashtra; Dr. R.D. Bhagat, Principal, VIVA College, Virar (W) Maharashtra; Non Teaching Staff, Dept. Biological sciences ,VIVA College Virar (W) Maharashtra.

REFERENCES

[1] Barnard JM, J. Chem. Inf. Comput. Sci. 1991; 31:64–68.
[2] Dobson CM, Nature 2004; 432: 824–828.
[3] Gordeeva EV., Molchanova M. S, Zefirov N. S. 1990; Tetrahedron Comput. Methodol. 3, 389–415.
[4] Zhao Y, Truhlar DG. Theor. Chem. Acc. 2008; 120:215–241.
[5] Baskin II, Zhokhova NI, Palyulin VA, Zefirov AN, Zefirov NS. Dokl. Chem. Engl. Transl. 2009; 427:172–175.
[6] Breiman L, Friedman J, Stone CJ, Olshen RA, Classification, Regression Trees, Chapman & Hall/CRC, Wadsworth, CA. 1984;
[7] Zhang, Sukumar N, Breneman C, Tropsha A. J. Chem. Inf. Model. 2006; 46: 844–851.
[8] Laggner C, Wolber G, Kirchmair J, Schuster D, Langer T. in Chemoinformatics Approaches to Virtual Screening Eds: A. Varnek, A. Tropsha;, RSC Publisher, Cambridge 2008; pp. 76 – 101.
[9] Erhan D, L'Heureux PJ, Yue SY, Bengio Y. J. Chem. Inf. Model. 2006; 46: 626–635.
[10] Geppert H, Humrich J, Stumpfe D, Gaertner T, Bajorath J. J. Chem. Inf. Model. 2009; 49:767–779.
[11] Varnek A, Gaudin C, Marcou G, Baskin I, Pey AK, Tetko IV. J. Chem. Inf. Model. 2009; 49: 133–144.
[12] Oprea TI, Kurunczi L, Olah M, Simon Z. SAR QSAR Environ. Res. 2001; 12:75 –92.
[13] Schçlkopf B, Smola AJ Learning with Kernels: Support Vector Machines, Regularization, Optimization, Beyond, MIT Press, Cambridge, MA, USA. 2002;
[14] Berge C, Hypergraphs, Elsevier. Amsterdam. 1989.
[15] Todeschini R, Consonni V. H,book of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.
[16] Halberstam NM, Baskin II, Palyulin VA, Zefirov NS. Dokl. Chem. Engl. Transl. 2002; 384: 140–143.
[17] Livingstone DJ, Salt DW. Rev. Comput. Chem. 2005; 21:287–348.
[18] Faulon JL, Churchwell CJ, Visco DP.Jr. J. Chem. Inf. Comput. Sci. 2003; 43:721–734.
[19] Johnson AM, Maggiora GM. Concepts, Applications of Molecular Similarity, Wiley, New York. 1990.
[20] Maggiora GM. J. Chem. Inf. Model. 2006; 46:1535–1535.
[21] Ruiz L, Garcia CG, Gomez-Nieto MA.J. Chem. Inf. Model. 2005; 45:1178–1194.
[22] Willett P, Barnard JM, Downs GM. J. Chem. Inf. Comput. Sci. 1998; 38:983–996.
[23] Besalu E, Girones X, Amat L, Carbo-Dorca R. Acc. Chem. Res. 2002; 35:289–295.
[24] Fradera X, Amat L, Besalu E, Carbo-Dorca R. Quant. Struct.-Act. Rel. 1997; 16:25–32.

[25] Kearsley SK, Smith GM. Tetrahedron Comput. Methodol. 1990; 3:615–633.

[26] Von Korff M, Steger. J. Chem. Inf. Comput. Sci., 2004; 44:1137–1147.

[27] Hristozov D, Oprea TI, Gasteiger J. J. Chem. Inf. Model., 2007; 47:2044–2062.

[28] Tetko IV. J. Chem. Inf. Comput. Sci. 2002; 42:717–728.

[29] Kristensen TG. J. Math. Chem. 2010; 48:287–289.

[30] Bemis GW, Murcko MA. J. Med. Chem. 1996; 39:2887–2893.

[31] Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Casaulta M, Odermatt A, Ertl P, Waldmann H, Proc. Natl. Acad. Sci. USA, 102:17272 –17277.

[32] Wetzel S, Klein K, Renner S, Rauh D, Oprea TI, Mutzel P, Waldmann H.. Nature Chem. Biol. 2009; 5:581–583.

[33] Renner S, Van Otterlo WAL, Seoane MD, Mçcklinghoff S, Hofmann B, Wetzel S, Schuffenhauer A, Ertl P, Oprea TI, Steinhilber D, Brunsveld L, Rauh D, Waldmann H.. Nature Chem. Biol. 2009; 5:585–592.

[34] Peltason L, Hu Y, Bajorath J. Chem Med Chem. 2009; 4:1864– 1873.

[35] Agrafiotis DK, Shemanarev M, Connolly PJ, Farnum M, Lobanov VS. J. Med. Chem. 2007; 50:5926–5937.

[36] Wester MJ, Pollock SN, Coutsias EA, Allu TK, Muresan S. Oprea T. I. J. Chem. Inf. Model. 2008; 48: 1311–1324.

[37] Radchenko EV, Palyulin VA, Zefirov NS. Chemoinformatics Approaches to Virtual Screening Eds: A. Varnek, A. Tropsha, RSC, pp. 150–181.

[38] Magee PS. QSAR: Rational Approaches to the Design of Bioactive Compounds Eds: C. Silipo, A. Vittoria;, Elsevier, Amsterdam,. 1991.

[39] Jolliffe IT. Principal Component Analysis, 2nd ed., Springer, Heidelberg. 2002.

[40] Larsson J, Gottfries J, Muresan S, Backlund A. J. Nat. Prod. 2007; 70:789–794.

[41] Mazzatorta P, Vracko M, Jezierska A, Benfenati E. J. Chem. Inf. Comput. Sci. 2003; 43:485–492.

[42] Wang YH, Li Y, Yang SL, Yang L. J. Chem. Inf. Model. 2005; 45:750–757.

[43] Tropsha A., Fourches D. Chem. Central J. 2009; 3.

[44] Bonchev D, Rouvray DH. Chemical Graph Theory. Introduction, Fundamentals, Gordon, Breach, New York. 1991; p. 300.

[45] Cherkassky V, Mulier F. Learning from Data: Concept, Theory, Methods. 2nd ed., Wiley, Hoboken, New Jersey, 2007.

[46] Zupan J, Gasteiger J. Neural Networks in Chemistry, Wiley- VCH, Weinheim. 1999.

[47] Baskin II, Gordeeva EV, Devdariani RO, Zefirov NS, Palyulin VA, Stankevich MI. Dokl. Akad. Nauk. SSSR. 1989; 307:613–617.

[48] Rissanen J. Ann. Stat. 1983; 11:416–431.

[49] Burden FR, Winkler DA. J. Med. Chem. 1999; 42:3183– 3187.

[50] Obrezanova O, Csanyi G, Gola JMR, Segall MD. J. Chem. Inf. Model. 2007; 47:1847–1857.

[51] Abdo A, Chen B, Mueller C, Salim N, Willett P. J. Chem. Inf. Model. 2010; 50:1012–1020.

[52] Breiman L. Mach. Learn. 2001; 45:5 –32.

[53] Agarwal S, Dugar D, Sengupta S. J. Chem. Inf. Model. 2010; 50:716–731.

[54] Joachims T, Hofmann T, Yue Y, Yu C. N. Commun. ACM 2009; 52:97–104.

[55] Fechner N, Jahn A, Hinselmann G, Zell A. J. Chemoinformatics 2010; 2.

[56] Nikolova N, Jaworska J. QSAR Comb. Sci. 2003; 22:1006– 1026.

[57] Rupp M, Schneider G. Mol. Inf. 2010; 29:266–273

[58] Kohonen T. Self-Organizing Maps, Springer. 2001.

[59] Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H. J. Chem. Inf. Model. 2007; 47:47–58.

[60] Mercier C, Fabart V, Sobel Y, Dubois. JE. J. Med. Chem. 1991; 34:934–942.

[61] Kurunczi L, Olah M, Oprea TI, Bologa C, Simon Z. J. Chem. Inf. Comput. Sci. 2002; 42:841–846.

[62] Bishop CM, Pattern Recognition, Machine Learning, Springer, New York, 2006.

[63] Satoh H, Sacher O, Nakata T, Chen L, Gasteiger J, Funatsu KJ. Chem. Inf. Comput. Sci. 1998; 38:210–219.

[64] Hert J, Keiser MJ, Irwin JJ, Oprea TI, Shoichet BK. J. Chem. Inf. Model. 2008; 48:755–765.

[65] Wawer M, Peltason L, Weskamp N, Teckentrup A, Bajorath J. J. Med. Chem. 2008; 51:6075–6084.