**Name – Mehdi Askari**

**Roll no – 03**

**Class – MSc – Bioinformatics – II**

**Assignment No – 2**

**Paper – II**

**SOLUTION:**

Q1 . Explain: Chemoinformatics and its history

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. 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. 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.

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. 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 databases. On the other hand, for many problems the necessary information is not available. 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: From less than 1% of all compounds have their 3D structure or have their infrared spectra. It is asked that can is this data even enough to gain 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. Cheminformatics has mainly dealt with small molecules, whereas bioinformatics addresses genes, proteins, and other larger chemical compounds. Chemistry and Bioinformatics complements each other for bimolecular process, like structure and function of proteins, the binding of a ligand to its binding site, the conversion of a substrate within its enzyme receptor, and the catalysis of a biochemical reaction by an enzyme.

Q2.     Why is it required to study Chemoinformatics? Describe

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.

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

Data can be categorised into four types:

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

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

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

4) Graphical data – any structure or data point may have associated graphical information such as spectra or plots. In all cases the data may be experimental or computed and the molecules may be real or virtual. The Internet is increasingly used to distribute data and information in chemistry. Database is defined as a self‐describing collection of integrated records, mainly stored on hard disk or CD – ROM.

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

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

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

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

Q3.     Write a note on history of Chemoinformatics.

The first, and still the core, journal for the subject, the Journal of Chemical Documentation, started in 1961. Then the first book appeared in 1971. 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.

1. Structure Searching

A connection table is essentially a representation of the molecular graph. 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. This provided the basis for full structure searching. Then, methods were developed for substructure searching, for similarity searching, and for 3D structure searching.

1. 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.

1. 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 analysing 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.

1. Molecular Modeling

In the late sixties, R. Langridge and co-workers 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.

1. 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 ’s 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.

Q4.     Comment on Chemoinformatics Vs Cheminformatics

Cheminformatics also referred as Chemiinformatics/Chemical information/Chemical informatics has been recognised in recent years as a distinct discipline in computational molecular sciences. Cheminformatics is also known as interface science as it combines Physics, Chemistry, Biology, Mathematics, Biochemistry, Statistics and informatics. The primary focus of cheminformatics is to analyse/simulate/modelling chemical information which can represented either in 2D structure or in 3D structure. Industry sectors such as, agrochemicals, food and pharmaceutical are distinct areas where cheminformatics plays significant role in the recent history of molecular sciences. Cheminformatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information. According to F.K. Brown “ The use of information and technology and management has become a critical part of the drug discovery process. Cheminformatics is the mixing of information resources to transform data into information and information into knowledge which is collectively referred as inductive learning.

Cheminformatics (sometimes spelled as chemoinformatics or chemo-informatics) is a relatively new discipline. Actually, it has emerged from several older disciplines such as computational chemistry, computer chemistry, chemometrics, QSAR, chemical information, etc. The names identifying these older disciplines can be controversial, but they have been studied for many years. Cheminformatics involves the use of computer technologies to process chemical data. Initial activities in the field started with chemical document processing (the Journal of Chemical Documentation was published in 1961 by ACS. What differentiates chemical data processing from other data processing is that chemical data involves the requirement to work with chemical structures. This requirement necessitated the introduction of special approaches to represent, store and retrieve structures in a computer system. Another challenge faced by this new field was to establish clear relationships between structural patterns and activities or properties. One of the earliest cheminformatics studies involved chemical structure representations, such as structural descriptors.

Q5.     Describe application of Chemoinformatics in details (Any 6)

The range of applications of chemoinformatics is wide, any field of chemistry can profit from its methods. The following lists different areas of chemistry and indicates some typical applications of chemoinformatics.

1. Chemical Information

* Storage and retrieval of chemical structures and associated data to manage the flood of data.
* Data 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

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

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.

Q6.     Elaborate on types of learning approach used in Chemoinformatics.

Three major aspects of Cheminformatics are:

i) Information Acquisition, is a process of generating and collecting data empirically

(experimentation) or from theory (molecular simulation)

ii) Information Management deals with storage and retrieval of information.

iii) Information use, which includes Data Analysis, correlation, and application to problems in

the chemical and biochemical sciences

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, etc), whereas the fundamental data type in chemoinformatics is centered on small molecules. 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 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 analysing 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.

**Name – Mehdi Askari**

**Roll no – 03**

**Class – MSc – Bioinformatics – II**

**Assignment No – 3**

**Paper – II**

**SOLUTION:**

Q1.     What does Chemical Structure Representation describe in Chemoinformatics study?

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. 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.

Two Dimensional (2D) representations of molecules only tell about atoms, which are bonded together. It does not tell about steric & electronic parameters and atom positions in 3D space. Three Dimensional representations of molecules have following challenges: 1) Molecules can adopt more than one low energy conformation and 2) The number of accessible structures is very large. So, there is need to represent molecular structures in 3D. The data stored in a 3D database either comes from Experimental methods or Computational methods. Structures are needed to be included in computer readable form. Emphasis laid on computational representation of molecular structures and creation of structural databases. Three dimensional structures determined by x ‐ray crystallography known for about 300,000 organic compounds. Or the largest database of infrared spectra contains about 200,000 spectra. It is only 1% of all the available compounds. The question asked is that can enough knowledge be gained from the known data to make predictions for those cases where the required information is not available. This is another reason why we need informatics methods in chemistry.

Q2.     Explain Structure Searching methods in detail.

A connection table is essentially a representation of the molecular graph. 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. This provided the basis for full structure searching. Then, methods were developed for substructure searching, for similarity searching, and for 3D structure searching.

This involves searching a database for an exact match with a specified query structure. Then only an exact match to this structure would be returned by a search. 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 connection table is essentially a representation of the molecular graph. 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. 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 analysing 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, CH = 3 and CH1H2 = 3. This tells us that C is unique; hence, amongst the carbons, only CH and CH1H2 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. CH gets a score of 1 (C) + 1 (H) + 3 (CH1H2) = 5. CH1H2 gets a score of 3 (CH) + 1 (H1) + 1 (H2) = 5. H gets a score of 3. H1 and H2 also get scores of 3. Scores based on summing the atomic numbers of bound atoms are also computed: CH gets a score of 13, CH1H2 gets a score of 8 and the protons all score 6. This means that CH is distinct from CH1H2.

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 H1 and H2. Thus, H is distinct from H1 and H2.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. This provided the basis for full structure searching. Then, methods were developed for substructure searching, for similarity searching, and for 3D structure searching.

Q3.     Discuss Chemometrics in detail.

The development of the discipline chemometrics is strongly related to the use of computers in chemistry. Some analytical groups in the 1970s were already working with statistical and mathematical methods that are ascribed nowadays to chemometric methods. Those early investigations were connected to the use of mainframe computers. The notation chemometrics was introduced in 1972 by the Swede Svante Wold and the American Bruce R. Kowalski. The foundation of the International Chemometrics Society in 1974 led to the first description of this discipline. In the following years, several conference series were organized, for example, Computer Application in Analytics (COMPANA), Computer-Based Analytical Chemistry (COBAC), and Chemometrics in Analytical Chemistry (CAC).

An actual definition of chemometrics is the chemical discipline that uses mathematical and statistical methods, (a) to design or select optimal measurement procedures and experiments, and (b) to provide maximum chemical information by analysing chemical data.

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 analysing 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. 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.

Q4.     Comment :

1. Quantitative Structure Activity Relationship  (QSAR)

Quantitative structure–activity relationship (QSAR) models are quantitative regression methods that attempt to relate chemical structure to biological activity. Quantitative structure–activity relationship and related methods have been applied extensively in a wide range of scientific disciplines, including chemistry, biology, and toxicology. In both drug discovery and environmental toxicology, QSAR models are now regarded as a scientifically credible tool for predicting and classifying the biological activities of untested chemicals. As we enter the new millennium, QSAR has become inexorably embedded as an essential tool in the pharmaceutical industry, from lead discovery and optimization to lead development. For example, a growing trend is to use QSAR early in the drug discovery process as a screening and enrichment tool to eliminate from further development those chemicals lacking druglike properties or those chemicals predicted to elicit a toxic response. This developing scenario portends the spread of QSAR beyond the pharmaceutical industry to human and environmental regulatory authorities for use in toxicology. Computer hardware and software improvements have been enabling technologies in QSAR development during the past decade. Within the pharmaceutical industry alone, the enormous financial incentives to accelerate the drug discovery process and to improve the odds of success by enriching the drug pipeline with more effective and less toxic candidates are powerful driving forces that have led to improved QSAR approaches and associated software. The integration of QSAR modeling with recent advances in hardware and software for data storage and management has further stimulated its wider implementation. The algorithms used in QSAR software also improved markedly, particularly with respect to the large and growing pool of descriptors used to characterize molecular structure and properties.

The fundamental assumption of QSAR is that variations in the biological activity of a series of chemicals that target a common mechanism of action are correlated with variations in their structural, physical, and chemical properties. Since presumably these structurally related properties of a chemical can be determined by experimental or computational means much more efficiently than its biological activity using in vitro or in vivo approaches, a statistically validated QSAR model is capable of predicting the biological activity of a new chemical within the same series in lieu of the time-consuming and labour-intensive processes of chemical synthesis and biological evaluation. Applied judiciously, QSAR can save substantial amounts of time, money, and human resources.

Quantitative structure–activity relationship modeling generally involves three steps: (1) collect or, if possible, design a training set of chemicals; (2) choose descriptors that can properly relate chemical structure to biological activity; and (3) apply statistical methods that correlate changes in structure with changes in biological activity. Obtaining a good-quality QSAR model with the ability to predict activity of a chemical outside the training set depends on many factors in the approach and execution of each of the three steps.

Quality of data

Data should come from the same assay protocol, and care should be taken to avoid interlaboratory variability. Any bad data points will tend to corrupt the proper correlation of structure and activity. Rules of thumb for a good QSAR data set are that the dose–response relationship should be smooth, the potency (or affinity) should be reproducible, the activity range should span two or more orders of magnitude from the least active to the most active chemical in the series, the number of chemicals used to build the QSAR model should be sufficiently large to ensure statistical stability, the activities of the chemicals should be evenly distributed across the range of activity, and the chemicals selected for the training set should possess enough structural diversity to span the range of chemistry space associated with the biological activity under study.

Descriptor selection

Many types of chemical structure descriptors are available from commercial software. Obtaining a statistically robust model is very much dependent on how well the selected descriptors can encode the variation of activity with structure. The more that is known at the molecular level about the biological mechanism of action of the chemicals, the better the chemist is able to select among the wide variety and types of specific molecular descriptors. Commercially available molecular modeling programs often include statistical tools to help in evaluating which descriptors best encode structure– activity variation. Some of these tools include the genetic algorithm (GA) in its various incarnations, which employs the evolutionary rules of natural selection to select the optimal (fittest) subset of descriptors for a particular problem.

Statistical methods

It is also critical that the QSAR method selected to develop the structure–activity correlation be suitable. Although the relationship between a molecular descriptor and biological activity may be linear or nonlinear, it is still common practice today to use linear approaches such as multiple (or multivariate) linear regression (MLR) or partial least squares (PLS) regression to construct the QSAR model. For nonlinear modeling, the Polynomial Neural Network (PNN) offers an alternative that combines the best features of Artificial Neural Networks (ANNs) and MLR/PLS by providing the inherent nonlinearity of the ANN with the desired analytical regression equation furnished by MLR and PLS. The most common scenario encountered in practice is for the number of possible descriptors to exceed the number of chemicals, a situation that can lead to chance correlations. Fortunately, soft modeling methods such as PLS reduce the risk of encountering chance correlations by transforming the dimensionality of the regression problem from chemical-descriptor space to so-called principal components (PCs) space.

1. Quantitative Structure Property Relationship (QSPR)

Quantitative structure–property relationships (QSPR) remain the focus of many studies aimed at the modeling and prediction of physicochemical and biological properties of molecules. A powerful tool to help in this task is chemometrics, which uses statistical and mathematical methods to extract maximum information from a data set.

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.

QSPR uses chemometric methods to describe how a given physicochemical property varies as a function of molecular descriptors describing the chemical structure of the molecule. Thus, it is possible to replace costly biological tests or experiments of a given physicochemical property (especially when involving hazardous and toxically risky materials or unstable compounds) with calculated descriptors, which can in turn be used to predict the responses of interest for new compounds. Chemometrics has provided new insight into the philosophy and theory behind QSPR modeling. It has been used to estimate properties such as density, boiling point, solubility, n-octanol–water partition coefficient, Henry’s law constant and vapor pressure of chemicals. QSPR has received significant contributions from various research schools. Various quantitative structure–property relationship (QSPR) models have been proposed for estimating the properties of a series of aliphatic alcohols.

The basic strategy of QSPR is to find an optimum quantitative relationship, which can be used for the prediction of the properties of compounds, including those unmeasured. It is obvious that the performance of QSPR model mostly depends on the parameters used to describe the molecular structure. Many efforts have been made to develop alternative molecular descriptors which can be derived using only the information encoded in the chemical structure. Much attention has been concentrated on “topological indices” derived from the connectivity and composition of a molecule which have made significant contributions in QSPR studies. Topological index has advantages of simplicity and quick speed of computation and so attracts the attention of scientists. Topological descriptors can explain most of the property modeled, as shown by some researchers.

1. 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 ’s to develop systems for computer-assisted structure elucidation (CASE) as a field of exercise for artificial intelligence techniques.

Computer-Assisted Structure Elucidation (CASE) has come to be a broadly accepted method for the derivation of novel or difficult chemical structures, especially those of natural products or drug metabolites and impurities. For completely error-free structure determination, one must first evaluate all isomers that truly match connectivity criteria defined by experiment – but it is impossible to do this manually without bias.

Modern NMR techniques give ever-increasing amounts of valuable data to substantiate structural evaluations, including lower limits of detection, and more information on connectivity’s. However, errors in published structures are rampant giving weight to the argument for computer-assisted evaluation of structures.

CASE follows a simple set of steps to comprehensively evaluate structures against available data:

1. Interpret experimental data to extract knowledge:

• Molecular Formula

• Integrals & Chemical shifts

• Multiplicities

• Connectivity’s

• Known fragments & exclusions

2. Search structure space to derive all possible structures

3. Rank-order based on set criteria

• Predicted chemical shift

Computer-Aided Structure Elucidation is an area of chemoinformatics and analytical chemistry and has been developed over a period of forty years. This development path has forced the developers of CASE systems to overcome many obstacles hindering the development of a software application capable of drastically reducing the time and effort required to determine the structures of newly isolated organic compounds. Large complex molecules of up to 100 or more skeletal atoms with topological peculiarity can be quickly identified using, for example, the expert system Structure Elucidator based on spectral data. Logical analysis of 2D NMR data frequently allows for the detection of the presence of COSY and HMBC correlations of "nonstandard" length and provides a solution to the problem. Fuzzy structure generation provides a possibility to obtain the correct solution even in those cases when an unknown number of nonstandard correlations of unknown length are present in the spectra. The relative stereochemistry of big rigid molecules containing many stereocenters can be determined using the StrucEluc system and NOESY/ROESY 2D NMR data for this purpose.

1. 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 must 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 co-workers, Hendrickson and Gelernter reported on their work on CASD systems. Later also at Toyohashi work on a CASD system was initiated.