# Assignment 1 (Chemoinformatics)

**Q1. Explain: Chemoinformatics and its history**

- "The application of management and information technology has evolved into a crucial step in the drug discovery process. Chemoinformatics is the blending of those information resources with the goal of turning data into knowledge and knowledge into information with the intention of helping people make better judgments more quickly in the domain of drug lead organisation and identification.
- Dr. Brown defined chemoinformatics in the manner stated above in 1998.
- Chemoinformatics's current application goes beyond simple drug discovery.
-
- Although the name "chemoinformatics" has been around for more than 20 years, the field's roots go back far further.
- The Journal of Chemical Documentation (JCD), which later changed its name to Journal of Chemical Information and Modelling, was founded in 1961 and is the oldest and most significant journal in this discipline (JCICS).
- Chemoinformatics is a field of IT that uses computers to facilitates the following for large quantities of chemical data:
    1. Collection
    2. Storage
    3. Analysis
    4. Manipulation
- This chemical data includes:
    1. Chemical formulas
    2. Chemical structures
    3. Chemical properties
    4. Chemical spectra
    5. Biochemical and Biological activities
- The most common uses of chemoinformatics are:
    1. Virtual screening
        a. This involves the creation of large *in silico* virtual libraries of compounds, which are then submitted to a docking programming in order to identify the most active members
    2. Quantitative Structure-Activity Relationship (QSAR)
        a. This is used to predict the activity of compounds from their structures.

**Q2. Why is it required to study chemoinformatics? Describe.**

- Using the right database, chemoinformatics is essential for maintaining and providing access to the vast amount of chemical data created by chemists.
- In order to represent intricate links between the structure of a chemical molecule and biological activity, or the influence of reaction conditions or chemical reactivity, the study of chemistry also required a new technique for information extraction from data.
- A significant amount of chemical data is being produced by recent chemical breakthroughs for drug discovery. (Explosion of information)
- This has led to a necessity for sophisticated drug research and development processes to effectively gather, organise, analyse, and utilise chemical knowledge.
- Even though the name "chemoinformatics" wasn't developed until much later, in 1998, the relationship between chemistry and information management began to intersect in the middle of the 1970s.The modern drug discovery and development pipeline process, starts with disease selection, target identification, lead identification, lead optimization, pre-clinical trial testing, clinical trial testing, approval and circulation.

- This takes a lot of time to do *ex-silicon* hence the involvement of computers (Chemoinformatics) in the early stages speeds up and reduces cost of drug design.
- Currently, Target identification, Lead identification, Lead optimization is done *in-silico.*

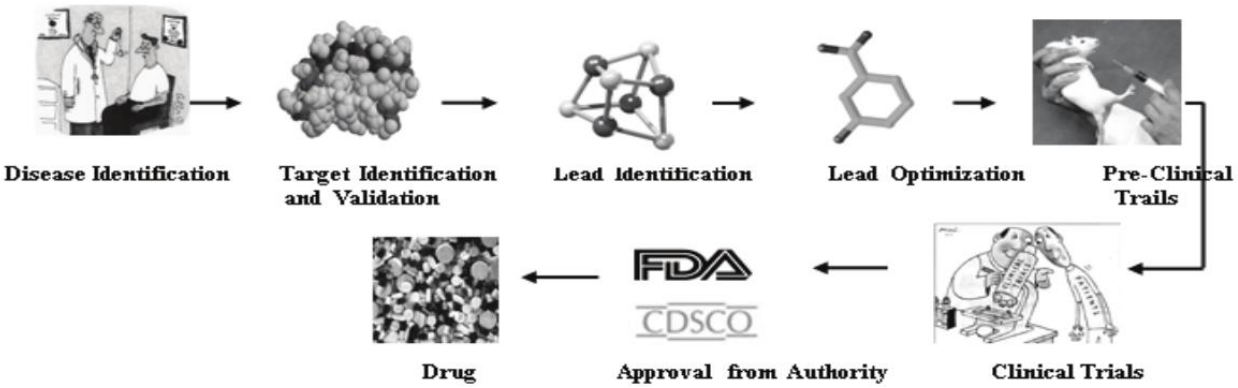


Fig: Modern Drug designing process

**Q3. Write a note on history of chemoinformatics**

- Chemoinformatics has been an active field in various guises since the 1970s and earlier, with activity in academic departments and commercial pharmaceutical research and development departments.
- The term chemoinformatics was defined in its application to drug discovery by Dr. F. K. Brown in 1998
- "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 optimization".
- Since then both spellings "Cheminformatics" and "Chemoinformatics" has been used interchangeably.
- In 1946 King et al, publicised an article illustrating the use of IBM's business accounting machines in carrying out the construction of the rotational spectra of asymmetric rotors by the evaluation of mathematical equations for line position and line intensity. This is regarded as the first work involving computer technology in chemistry and the birth of Chemoinformatics.
- In 1957 IBM made the first transistor-based computers as well as FORTRAN, a high-level programming language which made computers a generally available commodity for scientists.
- 1951 marked the invention of the first database system developed by Chemical abstract service (CAS) of the American Chemical Society.
- The following years from 1950s to 1990s saw exponential growth in both the capability of computers as well as the data available about chemistry.
- By the end of 1990s advancements in chemoinformatics had led to the invention and use of Database Systems, Case expert Systems, CASD expert systems and 3D structure builders.
- This was the turning point for chemoinformatics, which then led to the cascade of events that led to the current state of chemoinformatics / Drug design.

**Q4. Comment on Chemoinformatics vs Cheminformatics**

- Cheminformatics and Chemoinformatics have both been used interchangeably in research papers.
- Cheminformatics is being used more frequently (2.5 times more) than chemoinformatics since 2009.
- 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 detail**

- Applications of Cheminformatics are:
  1. Strong data generated through experiments or from molecular simulation Retrieval of chemical structure from chemical database
  2. Prediction of physical, chemical and biological properties of chemical compounds
  3. Elucidation of the structure of a compound based on spectroscopic data.
  4. Structure, substructure, similarity and diversity searching from chemical database
  5. High throughput screening (HTS) is the integration of technologies (lab automation, assay tech, micro plate based instrumentation
  6. Docking – Interaction between two macromolecules
  7. Drug Discovery
  8. Molecular Science, Materials Science, Food Science, Atmospheric chemistry, Polymer chemistry, Textile chemistry, Combinatorial organic synthesis

**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 modelling.