Prarthi Kothari
M.Sc. - Bioinformatics - Part - II
Roll No. - 115

Assignment 1

Q1] Explain : Chemoinformatics and its history.

Ans — **Definitions —**

(1) The set of computer algorithms and tools to store and analyze chemical data in the context of drug discovery.

(2) 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 to knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization.

(3) Greg Paris → "Chemoinformatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization and use of chemical information.

- **History of Chemoinformatics —**

- Problems in chemistry — modeling complex relationships, profusion of data, lack of necessary data.

(1) Early 60s → Explored various forms of machine readable chemical structure representations as → basis for building databases of chemical structures and reactions.

↳ Connection Tables — gained universal acceptance — represent molecules by lists of the atoms and of the bonds in a molecule.

(2) 2nd half of the 60s → Connection Tables used for the Chemical Abstracts Registry System.
↳ Problem → Graph isomorphism (defining from a set of potential representations of a molecule a single one as a unique one).
↳ First solution → Morgan algorithm — numbering the atoms of a molecule in a unique and unambiguous manner.
↳ Subsequent solutions developed — full structure searching, substructure searching, similarity searching, 3D structure searching.

(3) 1964 — Hansch and Fujita — showed that the influence of substituents on biological activity data can be quantified. (built on the work of Hammett and Taft in the 50s).

(4) Last 40 years — establishment of — Quantitative Structure - Activity Relationships (QSAR) and Quantitative Structure - Property Relationships (QSPR) as individual fields with their own journals, societies, and conferences.

(5) 70s → Introduced Pattern recognition methods to analyze chemical data.

(6) 90s → Prominence of Artificial Neural Networks for analyzing chemical data.

(7) Late 60s → (a) R. Langridge and coworkers — developed methods for visualizing 3D molecular models on the screens of Cathode Ray Tubes.

(b) C. Marshall — started visualizing protein structure on graphic screens.

(8) 1960 → Developed systems for computer-assisted structure elucidation (CASE) as a field of exercise for artificial intelligence techniques.

(9) 1964 → DENDRAL project — at Stanford University.

(10) Late 60s → Other CASE approaches initiated by —
(a) Sasaki at Toyohashi University of Technology.
(b) Munk at University of Arizona

(11) 1969 → (a) Corey and Hipke → development of a synthesis design system. (CASD — Computer-Assisted Synthesis Design).
(b) Ugi and coworkers
(c) Hendrickson
(d) Gelernter
} reported their work on CASD.

Q2] Why is it required to study Chemoinformatics ? Describe.

Ans – Chemoinformatics plays a key role to maintain and access enormous amount of chemical data, produced by chemists, by using a proper database.

- The field of chemistry needs a novel technique for knowledge extraction from data to model complex relationships between the structure of the chemical compound and biological activity or the influence of reaction condition on chemical reactivity.

- 3 major aspects of Cheminformatics are —

(1) Information Acquisition — process of generating and collecting data empirically (experimentation) or from theory (molecular simulation).
(2) Information Management — deals with storage and retrieval of information
(3) Information Use — includes Data Analysis, correlation and application to problems in the chemical and biochemical sciences.

- Areas that require Cheminformatics —
(1) Analysis and Modeling
(2) Environmental Effects and Hazards
(3) Spectroscopy
(4) Environment
(5) Pharmacology
(6) Regulations
(7) Toxicology
(8) Chemical and Physical Reference Data.

Q3] Write a note on the history of Chemoinformatics.

Ans -

	Period	Key Developments	Significance
(1)	Early Beginnings (1960s - 1970s)	- Initial use of computers for chemical data processing. - Journal of Chemical Documentation established (1961).	- Laid the groundwork for integrating computer science with chemical information.
(2)	Emergence of Key Concepts (1980s)	- Formalization of methodologies in structure representation and database retrieval. - Development of SMILES notation for efficient data storage and searching. - Introduction of structural descriptors for computational representation.	- Standardized methods for handling and analysing chemical information.
(3)	Formalization and Growth (1990s)	- Coining of the term "cheminformatics" by F.K. Brown (1998). - Integration in pharmaceutical R&D for drug discovery and optimization. - Development of QSAR models using statistical and machine learning techniques.	- Established cheminformatics as a crucial field in drug discovery and introduced advanced computational methods.

PTO →

	Period	Key Developments	Significance
(4)	Advancements in Computational Techniques (2000s)	- Increased computational power leading to enhanced capabilities. - Introduction of sophisticated algorithms and software tools. - Establishment of QSPR models and specialized journals like Journal of Cheminformatics (2009).	- Expanded possibilities for data mining, virtual screening and solidified the academic presence of the field.
(5)	Current State and Future Directions	- Integration with machine learning, data science and bioinformatics - Applications in drug discovery, materials science and environmental chemistry. - Expected advancements driven by AI and big data analysis for predicting chemical behaviours and optimizing processes.	- Chemoinformatics as a vital component of modern chemistry with future potential for further innovation.

Q4] Comment on Chemoinformatics v/s Cheminformatics.

Ans -

Category	Chemoinformatics	Cheminformatics
Definition	- A discipline that involves the extraction, processing and analysis of chemical data, particularly focusing on the relationships between chemical structures and their properties.	- A broader term that encompasses both informatics aspect and the computational chemistry methods used to analyze chemical data.
Involves	- Descriptor computations - Structural similarity assessments - Application of machine learning techniques to predict chemical and biological properties.	- Computational chemistry - QSAR - Chemometrics - to facilitate the understanding and prediction of chemical behaviours and interactions.
Focus	- Concentrate on data extraction and processing techniques.	- Stronger emphasis on theoretical modeling of chemical properties and behaviors.
Applications	- Drug discovery - focusing on optimizing compound selection, and predicting ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) properties.	- Materials science - Environmental chemistry - Drug discovery.

Q5] Describe application of Chemoinformatics in detail.

Ans – (1) Chemical Information —
- storage and retrieval of chemical structures and associated data to manage the flood of data
- dissemination of data on the internet.
- crosss 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.

(6) High Throughput Screening (HTS) —
- Integration of technologies (laboratory automation, assay technology, microplate-based instrumentation, etc.) to quickly screen chemical compounds in search of a desired activity.

Q6] Elaborate on Types of Learning Approach used in Chemoinformatics.

Ans - (1) Deductive Learning —

- based on established theories and principles.
- utilizes existing knowledge to make predictions about chemical properties or behaviors.
- Key characteristics —

(a) Theory - driven → begins with a general theory or model and applies it to specific cases.

(b) Model Validation → often requires validation against experimental data to ensure that the predictions align with observed outcomes.

(c) Application → QSAR modelling

(2) Inductive Learning —

- focuses on deiving general rules from specific exemples.
- Key features —

(a) Data - driven → analyzes large datasets to identify patterns and relationships without relying on pre-existing theories. It builds models on empirical data.

(b) Machine Learning Algorithms → Various machine learning techniques, such as Support Vector Machines (SVM), Random Forests and Artificial Neural Networks are employed to learn from data.

(c) Applications → High-throughput screening, drug discovery.

Prarthi Kothari
M.Sc. Bioinformatics – Part – II
Roll No. – 115

Assignment 2

① **Title** – "A molecular fragment cheminformatics roadmap for mesoscopic simulation"

② **Introduction** –

- Initially the significance of mesoscopic simulations in studying large molecular ensembles was discussed, which consists of millions of atoms.
- It highlights the limitations of traditional simulation methods, which often rely on quantum chemical calculations or molecular mechanics.
- The paper emphasizes the need for a chemoinformatics approach to enhance the accessibility and efficiency of mesoscopic simulations, specifically focussing on Molecular Fragment Dynamics (MFD).
- The authors aim to outline a comprehensive cheminformatics approach / roadmap that integrates various computational techniques to facilitate the practical application of MFD simulations.

③ **Methods** –

- The authors proposed a roadmap consisting of four building blocks essential for implementing the cheminformatics approach in MFD simulations–

(a) Fragment structure Representation –

↳ The development of a notation system called fSMILES is introduced, which allows for the representation of molecular fragments connected by harmonic springs.

↳ This notation is designed to be intuitive and similar to the established SMILES representation.

↓

(b) Operations on Fragment Structures –

↳ A library of functions is necessary for parsing and validating fSMILES strings, converting them into machine-readable formats

and mapping fragments to spatial coordinates for simulations.

↓

(c) Description of compartments — This involves defining compartments with specific compositions and structural alignments to organize the molecular fragments within the simulation environment.

↓

(d) Graphical Setup and Analysis —

↳ Tools for visualizing and analyzing the entire simulation box are discussed, enabling researchers to effectively interpret simulation results.

(4) Results —

- The results section elaborates on the four building blocks of the proposed cheminformatics roadmap.
- It details the implementation of the fSMILES notation for fragment representation and provides examples of its applications.
- The authors demonstrate how existing open-source cheminformatics software can be leveraged to meet the requirements of the roadmap.
- The section also discusses the potential for automating the conversion of peptides and proteins into fSMILES representations, which can facilitate simulations involving complex biomolecules.

(5) Conclusion and Discussion —

- The authors conclude that incorporating a cheminformatics layer into mesoscopic simulation techniques like MFD can significantly improve their practicality and broaden their applications in molecular sciences.

- This approach addresses the complexities and challenges associated with simulation setups, making it easier for researchers to utilize mesoscopic simulations effectively.
- The discussion emphasizes the importance of molecular fragment cheminformatics as a catalyst for advancing MRD and similar simulation techniques.

(6) References –

Trusković, A., Daniel, M., Kuhn, H., Neumann, S., Steinbeck, C., Zielesny, A., & Epple, M. (2014). A molecular fragment cheminformatics roadmap for mesoscopic simulation. Journal of cheminformatics, 6(1), 45.
https://doi.org/10.1186/s13321-014-0045-3

Prarthi Kothari
M.Sc. Bioinformatics — Part II
Roll No. - 115

Assignment - 3

Q1] What does Chemical structure Representation describe in Cheminformatics study ?

Ans - Cheminformatics involves storing, finding and analyzing molecular structures using the data-processing powers of computers to match chemical compounds with literature publications, measured properties, synthetic procedures, spectra and computational studies.

- Cheminformatics depends upon the use of representations of molecular structures and related data that are understandable both to human scientists and to machine algorithms.

① Ambiguous representations — refers to more than one chemical entity.
② Unambiguous representations — each name or formula refers to exactly one chemical entity.

- A chemical structure representation contains 2 kinds of information —
① Explicit → directly represented in a data structure.
② Implicit → interpreted on the basis of knowledge of general principles.

- Structural formula = any formula that indicates the connectivity of a compound → which of its atoms are linked to each other by covalent bonds.

- Connection table = the organized structural information defined in a molecular graph in a form that is easier to read and to order in a list.

- Some forms of machine-readable representations —
① graphic visualizations ② line notations
③ other descriptive forms such as nomenclature.

(1) Graphic visualizations —

(a) 2D coordinates — stored in connection table and can be used to infer and display chemical information, including the basic structural formula, E/Z geometry of alkene-like double bonds and the cis/trans isomerism of ligands in a square planar metal complex or substituents on a cyclic alkane.

(b) 3D (x, y, z) coordinates — used to display the conformation of a molecule; may be determined experimentally (typically via x-ray crystallography), or calculated (using force fields, quantum chemistry, molecular dynamics or composite models such as docking).

(2) Line Notations —

↳ Chemical structures represented as a linear string of symbolic characters that can be interpreted by systematic rule sets.

↳ Used to determine —

(a) whether molecules are the same

(b) how similar they are

(c) whether one molecular entity is a substructure of another

(d) whether two molecules are related by a specific transformation

(e) what happens when molecules are cut into pieces and grafted together at different positions.

↳ Examples of Line Notations —

(a) Kliswcosser Line-Formula Notation (WLN)

(b) Sybyl Line Notation (SLN)

(c) Representation of Structure Diagram Arranged Linearly (ROSDAL)

(d) Simplified Molecular-Input Line-Entry System (SMILES)

(e) IUPAC Chemical Identifier (InChI).

Q2] Explain structure searching Methods in detail.

Ans - (1) Exact structure search —

(a) Direct comparison — Linearly compares the query structure with each database entry. → Inefficient for large datasets.

(b) Hashing — Creates a hash value for each molecule, allowing for rapid lookup.

(c) Canonicalization — Ensures that different representation of the same molecule produce identical search results.

(2) Substructure search —

(a) Graph isomorphism — Determines if a smaller graph (substructure) is identical to a subgraph of a larger graph (molecule).

↳ Computationally expensive for large molecules.

(b) Fragment-based indexing — Breaks down molecules into fragments and indexes them. Faster than graph isomorphism but may miss some matches.

(c) Fingerprint-based substructure search — Uses fingerprints to approximate substructure searches. Faster but less accurate than graph-based methods.

(3) Similarity search —

(a) Fingerprint-based similarity — Compares the similarity of two molecules based on their fingerprints. Efficient but depends on the quality of the fingerprint.

(b) Shape-based similarity — Considers the 3D shape of the molecules, important for drug discovery where shape complementarity is crucial.

(c) Property-based similarity — Compares molecules based on physiochemical properties.

Q3] Discuss chemometrics in detail.

Ans – Chemometrics is the science of extracting information from chemical systems using mathematical and statistical methods.

- It is an interdisciplinary field that combines elements of chemistry, statistics, applied mathematics and computer science to address problems in various domains, including chemistry, biochemistry, medicine, biology and chemical engineering.

- Key aspects of chemometrics —

(1) Multivariate Calibration — aims to establish relationships between independent and dependent variables.

(2) Signal processing — a critical component, particularly the use of pretreatments to condition data prior to calibration or classification. Techniques like noise reduction and baseline correction are commonly employed.

(3) Performance Characterization — Chemometrics is quantitatively oriented, so considerable emphasis is placed on performance characterization, model selection, verification, validation and figures of merit.

- Applications —

- Chemometrics is heavily used in analytical chemistry, metabolomics, and process analytical technology.
- It enables the analysis of large, complex datasets to uncover hidden patterns and relationships, ultimately advancing the state of the art in analytical instrumentation and methodology.

Q4] Comment —

(1) Quantitative Structure Activity Relationship (QSAR) —

Ans - QSAR refers to a computational modelling approach that establishes a mathematical relationship between the chemical structure of compounds and their biological activity.

- QSAR models typically use a set of molecular descriptors — quantitative representations of molecular properties derived from chemical structures — to predict the activity of new compounds.

(2) Quantitative Structure Property Relationship (QSPR) —

Ans - QSPR is similar to QSAR but focuses on the relationship between chemical structures and their physiochemical properties rather than biological activities.

- QSPR models aim to predict properties such as solubility, boiling point or toxicity based on the molecular structure.

(3) Computer Assisted Structure Elucidation (CASE) —

Ans - CASE involves using computational tools to deduce the structure of unknown compounds based on experimental data, such as mass spectroscopy (MS) and nuclear magnetic resonance (NMR) spectroscopy.

(4) Computer Assisted Synthesis Design (CASD) —

Ans - CASD refers to the use of computational tools and algorithms to plan and optimize synthetic routes for chemical compounds. This method aids chemists in designing efficient and feasible synthetic pathways.