Cheminformatics: A Modern tool in Drug Discovery

Dr. Mamta Sharma[1] Er. Anil Kumar Dahiya[2] Dr. Sunita Dahiya[3]
[1]Principal [3]Assistant Professor
[2]Department of Chemical Engineering
[1,3]Aditi Mahavidyala (D.U) [2]I.E.I (India)

Abstract— The introduction of the high throughput screening and the combinatorial chemistry techniques has resulted in to a huge increase in the volumes of data about structures and their bioactivities. The explosion of data has increased the need for integration of chemical information with molecular modeling techniques. Knowledge management is playing a major role in almost all chemical and pharmaceutical companies. New cheminformatics units are created to assist on-going drug discovery programs. Many studies have appeared on cheminformatics; however this paper briefly outlines the process, methodology and effective implementation of cheminformatics in both small and large organizations successfully by bridging vital link between theoretical design and in drug design through extraction of information from the data and convert into knowledge. This research will provide valuable information on the Impact of cheminformatics on drug discovery cycle, comparison of old method and new method of drug discovery, understanding bioinformatics, chemogenomics, cheminformatics and molecular modeling, use of combitorial chemistry, property evaluation, protein and ligand selection process by HTS.

Ke words: Cheminformatics, Drug Discovery, HTS

I. INTRODUCTION

Competition and cost over the years has changed the drug design paradigm from the old hit and trial approach to a new automated drug design approach which allows generation of tailor made design of active molecules. It has not only resulted in targeted drug discovery but has also reduced drug development cycle time. The need for introducing newer molecules that are superior using automated approach will make drug discovery a highly knowledge specific.

Knowledge management is playing a major role in almost all chemical and pharmaceutical companies. New cheminformatics units are created to assist on-going drug discovery programs. Many studies have appeared on cheminformatics, however this paper briefly outlines the process, methodology and effective implementation of cheminformatics in both small and large organizations successfully by bridging vital link between theoretical design and in drug design through extraction of information from the data and convert into knowledge.

This study is intended to answer the two major questions in the field of drug designing-

- What to test next?
- What to make next?

II. THE IMPACT OF CHEMINFORMATICS ON DRUG DISCOVERY CYCLE

With the introduction of first tools in 1995 now currently more than 20 tools are available. From open, modular, platform to vendor independent architecture with integration with other scientific applications it has more than 1'800 registered users and over 5'000 jobs submitted each month while 20 million molecules being processed per year on an average.

Typical Cheminformatics activities at pharma industry include molecular databases, large-scale data analysis, knowledge discovery calculation of molecular properties / descriptors, estimation of ADME characteristics, toxicity alerting, navigation in chemistry space, virtual screening and support for HTS –hit list triaging, support for combinatorial chemistry and molecule optimization.

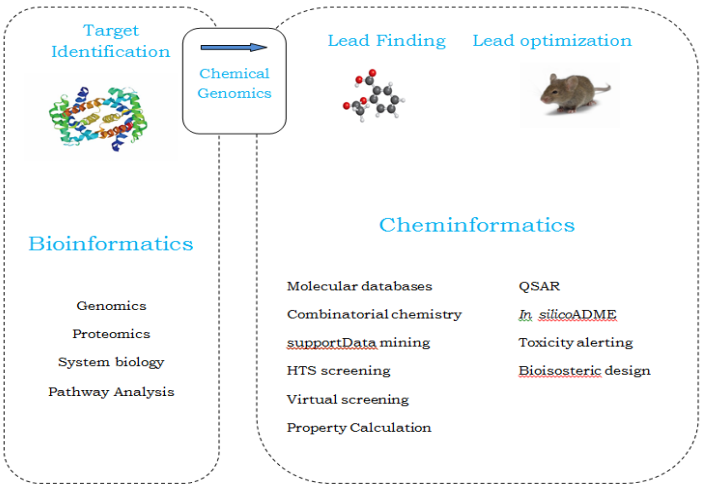

Fig. 1: Cheminformatics on Drug Discovery Cycle

A. Molecular Database

Database in pharmaceutical companies holds millions of structures and related data, normalization of chemical structures (nitro, tautomers …), all data need to be validated and checked for correctness, interface must support user-friendly data mining and visualization of large datasets, responsiveness -substructure and similarity searches within seconds, chemically interpretable results -pharmacophore searches, pharmacophore fingerprints.

Avalon in-house database written in Java, containing all in-house and many reference structures, results of biological screens and many additional data. Allows efficient data-mining, reporting and SAR analysis.

B. Large-Scale Data Mining

Data Mining = Knowledge Discovery in Large Databases
Analyzing large amount of data to obtain useful information (in a form of pattern, rule, cluster …) leading to understanding of relationships within data and correct decisions.
Data mining techniques used in cheminformatics:

- Classical QSAR, regression analysis
- Bayesian statistics
- Clustering
- Neural networks
- Decision trees

C. Self-Organizing Neural Networks

Self-organizing (Kohonen) NN is a mathematical tool used to simplify complex multidimensional data by reducing their dimensionality, allowing thus visual processing. Processed data are expressed as a 2-dimensional map.

D. Classification of GPCR Ligands:

Identification of properties and structural features typical for GPCR ligands are done through self-organizing neutral networks.

E. Calculation of Molecular Properties:

- Properties need to be calculated for datasets containing ~106 molecules (in-house data, virtual libraries, catalogues).
- Calculations need to be fast.
- Descriptors should be interpretable and physically meaningful.
- Properties should cover all important types of protein-ligand interactions

Currently the most useful global properties are logP, MW, PSA(polar surface area), HBD and HBA counts, number of rotatable bonds. Many others are used, but they are less interpretable + highly inter correlated.

Properties relate directly to the biological effect of drugs and their fate in organism, and are therefore frequently needed in medicinal chemistry. We need to take in account water solubility, pKa -acidity / basicity estimation, drug transport characteristics, intestinal absorption, blood-brain barrier penetration, caco-2 permeability, plasma-protein binding, efflux, toxic and metabolic characteristics.

F. Navigation in Chemistry Space

The chemistry space size of the known compounds or molecules serves as the bank for the organized database. Almost 35 million molecules are registered in CAS, 19 million compounds in PubChem, 36 million entries in the chemical structure lookup service, approximately 500,000 molecules with (known) biological activity exist for now. It is a very large number of possible (virtual) molecules. Chemistry space is multidimensional; to process or understand it, we need to characterize it and to reduce its dimensionality. Chemistry space may be characterized by: physicochemical global molecular properties (logP, PSA …) and sub structural features (fragments, fingerprints, pharmacophore...).

G. The Scaffold Tree –Basic Algorithm

Retain the molecular framework as classification element where Exocyclic and "exolinker" double bonds are part of the molecular framework. Instead of removing atom & bond type and ring size information prune less important rings one by one. Use prioritization rules to decide which ring to remove first. Use small and generic set of rules.

H. Virtual Screening

We can understand screening as selection of molecules having the highest probability to be active and to be developed to successful drugs from a large collection of screening samples or virtual molecules. In-house company archives contain 2-5 million molecules (in house synthesis, acquisitions, mergers, combichem libraries) and 20-30 million screening samples available commercially.

Selection criteria: reliable properties (solubility, stability, absence of too reactive fragments) drug-likeness, no toxicity or adverse effects, diversity or novelty and target focus.

I. Learning from the Nature

Natural products (NPs) have been optimized in a several billion year long natural selection process for optimal interaction with bio molecules.

J. High Throughput Screening - HTS

HTS is screening of >1 million molecules on many targets routinely in an automatic way.

The challenges for cheminformatics are to process screening results and identify hits, worth of further follow-up, lead identification, hit list triaging and support of new types of screening (high content screening, pathways).

- HTS Workflow.
- Run HTS, collect the data.
- Identify "active" compounds (based on % inhibition cut-off).
- Organize actives into groups (clustering, maximal substructure analysis, common scaffold)
- Visualize clusters of actives.
- Analyze inactive to identify those related to active series.
- Selected actives (primary hits) are further confirmed in dose/response assays to get EC50values, secondary assays and chemical validation to get validated hits.
- Use machine learning techniques to develop SAR models for validated hits.

K. Combinatorial Chemistry – Molecule Library Design

This technology was introduced in the early 90s and has advantages like speed & economics, the combination of scaffolds and Rgroups allows creation of very large number of molecules quickly in automatic manner.

L. Cheminformatics issues in Library Design:

How large should be combichem libraries? Which Rgroups and scaffolds to combine? Diverse libraries or targeted libraries. How to fill the "holes" in the chemistry space?

M. Early Combichem

Results of early combichem were quite a disappointment. Early combichem libraries were very large (100'000s molecules) molecules were large, hydrophobic, not diverse had low hit rates. This led to introduction of "drug likeness"–design of compounds with good physicochemical properties. Targeted libraries had design of smaller, more focused libraries when information about target is available (i.e. kinase libraries).Use diverse libraries covering broadly chemistry space when little information about target is available we preferable use "primary screening" libraries.

N. Library Design Strategies

We can classify library design in two basic design strategies:

1) Reactant-based:

Building blocks are selected based only on their properties not considering properties of products.

2) Product-based:

Monomers are selected based on the properties of final products. This approach is much more computationally demanding but is more effective.

Trends in modern CombiChem refers to smaller (1000s molecules), targeted libraries, multi objective optimization (Pareto optimization) to optimize different properties at the same time, coverage of chemical space, price, information from pharmacophore search or docking used in design, natural product-like libraries active tools available in the market.

III. STATEMENT OF THE PROBLEM

Discovering drugs to a disease is still a challenging task for medical researchers due to the complex structures of bio molecules which are responsible for disease such as AIDS, Cancer, Autism, Alzheimer etc. Design and development of new efficient anti-drugs for the disease without any side effects are becoming mandatory in the recent history of human life cycle due to changes in various factors which includes food habit, environmental and migration in human life style complexity of diseases and competition among pharma companies.

IV. ANALYSIS

The modern pharmaceutical 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 (Drug in market). In traditional drug discovery phase, the process which cost more time and money is replaced with lead identification and lead optimization process in modern drug discovery system. Each phase has an interaction component that transfers data, knowledge and information to one another.

Cheminformatics plays a key role to maintain and access enormous amount of chemical data, produced by chemist (more than 45 million chemical compounds are known and the number may increase in million every year,) by using a proper database. Also, 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 and the influence of reaction condition on chemical reactivity.

Average life span of human being is gradually decreasing in the recent medical history due to the higher influence of new diseases. Identifying and understanding structural and functional behavior of chemical compounds or bio molecules are one of the challenging issues for medical researchers. Cheminformatics is an emerging field which is used for better understanding of bio molecules. This paper primarily focuses on cheminformatics and its applications on drug discovery,

issues of traditional discovery and importance of modern drug discovery system. This in turn helps chemists and researchers for developing drugs without side effects.

V. SCOPE OF FUTURE STUDY

The future holds a very bright opportunity of reducing the time frame for animal and human clinical trials. A complete integration, from systems biology to virtual physiology, compounds will no longer be profiled at the molecular level, but also in terms of genetic and clinical effects. Among potentially novel tools, we anticipate machine learning models based on free text processing, an increased performance in environmental cheminformatics, significant decision-making support, as well as the emergence of robot scientists conducting automated drug discovery research. Furthermore, cheminformatics is anticipated to expand the frontiers of knowledge and evolve in an open-ended, extensible manner, allowing us to explore multiple research scenarios in order to avoid epistemological ‘‘local information minimum trap’’.

REFERENCES

[1] Brown FK: Cheminformatics: what is it and how does it impact drug discovery. Annu Rep Med Chem 1998, (33) 375–384.
[2] Blake JF: Cheminformatics – predicting the physicochemical properties of drug-like molecules. Current Opinion in Biotechnology 2000, (11) 104-107
[3] Willett P: Cheminformatics – similarity and diversity in chemical libraries. Current Opinion in Biotechnology 2000, 11:85-88