Journal of Pharmacology and Biomedicine ISSN 2456-8244

EDUCATIONAL ARTICLE

Drug Discovery Tools: Computer Aided Molecular Design and Chemo-Informatics

Richa Mishra and Brijeshkunvar Mishra *

Technocrats Institute of Technology-Pharmacy, Bhopal, Madhya Pradesh, India.

* Corresponding Author

Email: bjmishra08@gmail.com

ABSTRACT

The article reviews the different approaches used in computer aided molecular design. It also reviews current achievements in the field of chemo-informatics and their impact on modern drug discovery processes. The main data mining approaches used in chemo-informatics such as structural similarity matrices, and classification algorithms, are also outlined. In the conclusion, future prospects of chemo-informatics are also mentioned.

Keywords: Computer aided molecular design, chemo-informatics, drug discovery, combinatorial chemistry.

Introduction

The society today faces many challenges such as AIDS, bacterial resistance etc that can have a chemical solution. For this, there is a need to develop environmentally benign synthetic methods. A tremendous work has been done and a bewildering variety of structural motifs have been discovered in nature.

Traditional drug discovery process

There are seven sequential steps (Augen, 2002) in the drug discovery process: disease selection, target hypothesis, lead compound identification (screening), lead optimization, pre-clinical trials & clinical trials and pharmacogenomic optimization. If any of these steps is slow, it slows down the entire process. The slow steps are known as the bottlenecks. Earlier, the main bottlenecks were the time and cost of making and testing new chemical entities (NCE). To reduce the costs, new technologies were required to be found which could replace the older tedious approach of synthesis and screening of the NCE. With the invention of HTS and newer technologies, the use of robotics was introduced into screening. Through this process, hundreds of thousands of individual compounds can be screened with rapid pace (Gallop et al., 1994; Hetch, 2002). Chemical information technology helps us to appreciate the richness and variety of chemical structural complexity. Computer Aided Molecular Design, CAMD, is a combination of computational chemistry and information technology tools that helps us to discover new and useful compounds.

Computer Aided Molecular Design

CAMD is a unifying theme that focuses on why do we do chemistry and how do we decide what to synthesize and study. Chemistry emphasizes the development of predictive tools for understanding structure-function relationships, and the use of CAMD techniques enhances our ability to predict chemical reactivity and design useful compounds. The goal of CAMD is to find ligands that are predicted to interact strongly with a host. Alternatively, this procedure can be reversed to search for hosts that will interact strongly with a given ligand. CAMD is an outgrowth of rational drug design (Martin, 1991) where the interactions are protein or DNA binding with substrates. But CAMD is not restricted to drug design. As organic and physical chemists search for guest-host systems with specificity in binding and catalysis (Julius, 1991; Organic 'Tectons', 1995), the basic concepts of molecular field analysis and receptor mapping will be a unifying tool. Rapid advancements in chemistry will increasingly require an interdisciplinary approach; biochemistry, molecular biology, microbiology, cell biology, developmental biology will be key players along with the traditional areas of chemistry.

The ready availability of chemical structure databases is playing an important role in

enhancing the drug discovery approach and CAMD.

The basic principles of CAMD are outlined as shown in Fig. 1(Balbes, 1994).

CAMD can be done in two ways: ligand based or receptor based. Receptor based design stats with a known receptor, such as a protein binding site or supramolecular host. Ligand based design uses a known set of ligands, but an unknown receptor site. Both approaches are actually very similar.

Receptor based CAMD

The first phase is to determine the structure of the binding site using standard structural analysis from X-ray diffraction, NMR, or calculations involving molecular orbital or molecular mechanics and dynamics techniques. In the absence of structural information, homology of the known receptor sequence with known structures that have been identified through database search may be a good starting point.

The next phase is to generate a query for database searching. Building a simplified model of the receptor site generates the query. This model may be based on pharmacophore, which identifies a few specific interactions that are responsible for the binding. The pharmacophore can be generated by visual inspection or by computational techniques. In docking-based searches, the model is based on an analysis of steric interactions over the receptor site. Typically, a solvent accessible surface map is generated and binding pockets are identified on the host surface. The next phase is to search databases for ligands that may bind to the chosen receptor. The3D-pharmacophore is used in conformationally flexible searches for ligands that match the spatial distribution of the receptor or the receptor pocket can be used with auto-docking to find ligands that acoid close-contacts.

The results of the database search may be used directly or modified to produce candidates for further study. The new ligands or the hosts are then assessed for the use at hand. This assessment first involves docking the new molecule and evaluation of the full interaction by molecular orbital or molecular mechanics. Next, calculations are done to predict the binding constant or activity of the compound by using Gibb's free energy perturbation studies based on either Monte Carlo or Molecular dynamics simulations.

Ligand based CAMD

Ligand based design starts with a group of ligands that have known binding constants or biological activities. The first phase is to determine the structure of the ligands using standard structural analysis from X-ray diffraction, NMR, or calculations involving molecular orbital or molecular mechanics and dynamics techniques.

The next phase is to generate a query for database searching. Building a simplified model of the receptor site generates the query. This

model is based on a pharmacophore, as in receptor-based design. The pharmacophore can be generated by visual inspection or by statistical techniques. One popular statistical technique is 3D-QSAR as represented by the CoMFA approach (Cramer et al., 1988). 3D-QSAR maps the steric, charge, and hydrogen bonding interactions into a 3D-grid for each known ligand. These maps are then converted to find features that the active compounds have in common. The map of common features is then converted into a pharmacophore.

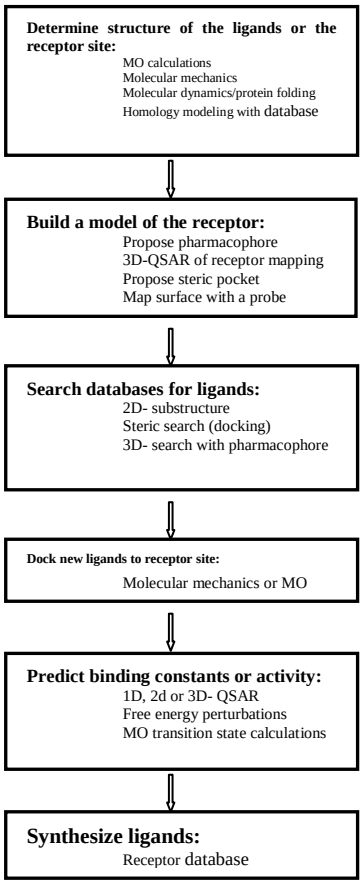

Fig. 1 Basic phases of CAMD

The next phase is to search databases for new ligands that may also bind to the chosen receptor. 2D-substructure search databases based on the known ligands can be used, but they have not been very useful. Instead, the 3D-pharmacophore is used in conformationally flexible searches for ligands that match the spatial distribution of the known ligands.

In both the techniques of CAMD, the candidates are synthesized and tested in the laboratory. Synthetic chemists increasingly use reaction database searches and artificial intelligence tools to design synthetic procedures.

The CAMD can be used to search databases and help synthesize thousands of compounds every day. But can chemists make thousands of compounds a day? Yes, the answer lies in a newer technique of synthesis called as the combinatorial chemistry.

Combinatorial chemistry

Chemists now-a-days use combinatorial chemical techniques to produce more new compounds in shorter time period. Combinatorial chemistry systematically and repetitively yields a large array of compounds from a set of different types of reagents called as "building blocks". But this process alone was not able to accelerate the drug discovery process. Therefore it was proposed that increasing the chemical diversity of compound libraries would enhance the drug discovery process. Chemo-informatics approaches would

now be introduced in order to optimize the chemical diversity of libraries.

Chemo-informatics and chemical diversity

To make a compound library with great chemical diversity, a variety of structural processing technologies for diversity analyses are created and applied. These computational approaches such as structural descriptor computations, structural similarity algorithms, classification algorithms, diversified compound selection and library enumerations are the components of chemo-informatics.

Technologies have been developed to recognize drug-like compounds from a diverse compound library (Ajay et al., 1998; Sadowski et al., 1998; Lipinski et al., 2001; Matter et al., 2001; Xu et al., 2000). These technologies have partly solved the screening problems. The technologies should be filtered in order to recognize *lead-like compounds* instead of *drug-like compounds*. The absorption, distribution, metabolism, excretion and toxicity (ADMET) parameters should be simultaneously optimized.

The lead optimization remains the most important bottleneck in the drug discovery process.

To improve the probability of a HTS being having efficient ADMET parameters has led to the use of chemo-informatics methods while generating data using high throughput technologies so as to assure better ADMET properties while the compound is still in the developmental phase. This approach is known as a *multi-parametric optimization strategy* (Baxter et al., 2001).

Origin of chemo-informatics

Chemo-informatics has emerged from several older disciplines such as computational chemistry, computer chemistry, chemo-metrics, QSAR, chemical information etc. Chemo-informatics involves the use of computer technologies to process chemical data which involves the working with chemical structures which in turn leads to necessity of introduction of special approaches to represent, store and retrieve structures in a computer system.

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 be easily searched by a computer. Finally, an atom-by-atom search algorithm can be applied to a smaller number of compounds.

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 is the Wiswesser Line Notation (WLN). In mid 70s it was considered as the best tool to represent, retrieve and print chemical structures (www.asis.org). 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 for storing and searching chemical structures in low performance computer systems. Later, a new linear notation called as SMILES™ was developed (Weininger, 1988; www.esc.syrres.com). to successfully represent a structure, a linear notation should be canonicalized. That is, one structure should not correspond to more than one linear notation string, and conversely, one linear string should only be interpreted as one structure.

Canonicalization

If a structure corresponds to a unique WLN or a unique SMILES™ string, then the structure search results in a string match. WLN was unable to meet this requirement often. The SMILES™ approach could do this only after canonical processing. A molecular graph (2D structure) can also be canonicalized into a real number through a mathematical algorithm. The real number is identified as a molecular topological index. However, two different structures can have the same topological index. Weiner was the first to report the molecular topological index (Wiener, 1947).

Dimension reduction and descriptor selection.

To view a one hundred- dimensional space, it is required to project the higher dimensional data space to two- or three- dimensional space. This is known as *dimension reduction.*

Mathematically, a library with **n** compounds and represented by **m (m**>3) descriptors is an **n x m** dimensional matrix. Many dimension reduction approaches are available.

Multidimensional scaling

Multidimensional scaling (MDS) (Cox and Cox, 2000) or artificial neural networks (ANN) methods are traditional approaches for dimension reduction. MDS is a non-linear mapping approach. In this, the objects are moved around in the space defined by the specified number of dimensions and, then checks how well the distances between objects can be reproduced by the new configuration. In other words, MDS uses a function minimization algorithm that evaluates different configurations with the goal of maximizing the goodness-of-fit (or minimizing "lack of fit") (www.statsoft.com).

Self-organizing map (SOM)

Self-organizing map (SOM) is one of the ANN methods. It is a vector quantization algorithm that creates reference vectors in a high-dimensional input space and uses them, in an ordered fashion, to approximate the input patterns in image space. To do this it defines local order relationships between the reference vectors so that they are made to depend on each other as though their neighboring values would lie along a hypothetical "elastic

surface"(Gasteiger and Zupan, 1993). The SOM is therefore able to approximate the point density function, p (x), of a complex high dimensional input space, down to a two-dimensional space, by preserving the local features of the input data (Bernard et al., 1998). ***PCA and FA***

Principal component analysis (PCA) (Joliffe, 1986) and factor analysis (FA) (Malinowski and Howrey, 1980) are usually used to filter out superfluous descriptors and, eliminate descriptors having minor information contributiwon. PCA is used to transform a number of potentially correlated variables (descriptors) into a number of relatively independent variables that then can be ranked based upon their contributions for explaining the whole data set. The transformed variables that can explain most of the information in the data are called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding components accounts for as much as the remaining variability as possible. The components having minor contribution to the data set may be discarded without losing too much information. FA uses an estimate of common variance among the original variables in order to generate the factor solution. A factor is a linear combination of original variables. The number of factors will always be less than the number of original variables. So, selecting the number of factors to keep for further analysis using common factor analysis is more problematic than is selecting the principal components. If the number of principal components or factors is less than four, then the multi-dimensional data can be graphed into two- or three- dimensional space. To validate the dimension reduction results, a technology called as the chemical structure-related data visualization is used.

Descriptor selection

Successful data mining depends on good descriptor selection. Correlation analysis can be used to understand the computational problem that is being tried to solve. The criteria used for selecting descriptors should be: (1) the selected descriptors should be informative, (2) the selected descriptors should be bioactivity related, (3) the selected descriptors should be independent of each other, (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 between different categories. Thorough research has led to the conclusion that the 2D descriptors perform significantly better than the 3D descriptors (Brown et al., 1996).

Pattern recognition

The core technology of data mining is pattern recognition. In chemo-informatics, regression and classification are commonly used pattern

recognition technologies. Regression analysis is applied to the variables that have continuous values. Table 1 shows certain common patterns in chemo-informatics (Xue et al., 2001).

In order to compare patterns, one needs similarity or distance measurements.

Table 1 Common pattern in chemo-informatics.

Method	Methodolody
Fingerprint	This pattern has no human bias as it is generated systematically from an algorithm. This is a topological pattern and is used in HTS data mining.
Regression	Regression methods are the most traditional approaches for pattern recognition. These methods are based on the assumption that the variables are continous and the curve shapes are pre-defined. In multi-dimensional data where the curve patterns are not known, genetic algorithms may be applied to partially solve the problem.
Generic structure or Markush structure	This is a topological pattern used by chemists for many years. It is determined by experience. It is an efficient way to represent an unlimited number of compounds with the same scaffold. Additional restrictions can be applied to make the pattern more specific. It is suitable for lead optimization and hit-to lead efforts.
Three-dimensional pharmacophore	This pattern is derived, manually or computationally, from a three-dimensional molecular model. The pattern is based upon a physical model and binding mechanism. It is sensitive to conformation changes. Better results are obtained when supported by crystal or NMR structural data. It is suitable for lead optimization.
Hierarchial clustering	This approach assumes the objects have hierarchical characters. The methods require similarity or distance matrices. The approach may produce multiple answers for users to explain or with which to experiment.
Non-hierarchial clustering	The approach assumes the objects have nonhierarchical characters, and the number of clusters is known prior the computation. The method requires similarity or distance matrices. The approach may produce multiple answers for users to explain or with which to experiment.
Decision tree classification	This approach is applied when there are a great number of descriptors and, the descriptors have various value types and ranges.

Similarity or distance metrics.

Many pattern recognition techniques require distance or similarity measurements to quantitatively measure the distance or similarity of two objects. Euclidean distance, Mahalanobis distance and correlation coefficients are commonly used for distance measurement.

The Tanimoto coefficient is commonly used for similarity measurements of bit-strings of structural fingerprints.

Clustering

Cluster analysis (CA) encompasses a number of different classification algorithms. CA algorithms belong to two categories: hierarchial and non-hieracrchial (partitional) clustering (Willet, 1987).

Correct clustering results rely upon: (1) proper structure representation, (2) suitable data normalization, and (3) carefully selected cluster algorithms and proper parameter settings. Data normalization is the basis for comparing experiments within large series when experimental conditions may not be identical.

Normalization ensures that the experimental quality of the data is comparable and, sound mathematical algorithms have been employed. Normalization includes various options to standardize data and to adjust background levels and correct gradients. The commonly used normalization functions are linear normalization, Ratio normalization and Z-score normalization (Jarvis and Patrick, 1973). ***Partitioning***

Partitioning algorithms, such as, decision trees, are non-parametric approaches. It is difficult for regression or parametric classification approaches to work on heterogeneous types of data. The excessively large number of descriptors can make clustering computation feasible. Decision trees are introduced to solve these problems. One of the most popular decision tree techniques is recursive partitioning (RP).

Application in drug discovery

The CAMD approaches and chemo-informatics have been widely applied in the drug discovery process. The major applications are outlined below:

1. *Compound selection*

The main tasks for compound selection are: (1) to select and acquire compounds from external sources that will provide complementary diversity to existing libraries, (2) to select for screening, from a corporate compound pool, a subset that provides diversity representation, (3) to select reagents to make a combinatorial library which will maximize diversity, and (4) to select compounds, from available compound collections, that are similar to known ligands yet, with different and novel scaffolds. Diversity-based compound selection has been done using many classification approaches.

2. *Virtual library generation.*

A virtual library can be generated using a computational approach. The criteria for generating a general (not focused) virtual library are: (1) diversity, (2) ADMET properties, and (3) synthetic accessibility. There are a number of ways to generate a diverse virtual library. However, it is challenging to make a virtual library that meets the criteria set forth above in (2) and (3). Although work on this aspect has been reported [86-87], more investigation is required.

3. *Virtual screening of compounds.*

Virtual screening is actually one of the computational tools used to filter out unwanted compounds from physical libraries or *in silico* libraries. If the target structure is known, one of the structure-based virtual screening methods that can be used is high throughput docking [92-93]. If the target structure is unknown, but the ligands from the literature or, competitors are known, then, similarity approaches can be applied (Xu and Hagler, 2002). If neither target structure nor ligand structure is known, then SAR patterns can be derived from experimental screening data

by statistical approaches. Also, virtual screening is a great tool for the design of a combinatorial library with a given target.

4. *Structure activity relationship studies.*

The purpose of Sequential HTS is to maximize receptor ligand interaction information by using HTS and CC technologies, discover novel leads as soon as possible and, minimize HTS and library production costs (Fig. 2). Sequential HTS screens compounds iteratively for activity, analyzes the results and, selects a new set of compounds for next screening, based on what has been learned from the previous screens. The iteration ends when the desired, nano-molar, novel leads are identified. Compound selection is driven by rapid SAR analyses using recursive-partitioning techniques. Although there are not many publications on the subject, sequential HTS has been studied in many pharmaceutical companies under different terminologies, such as: recursive screening, and progressive screening.

5. *ADMET calculation of compounds by computational methods.*

Higher-throughput, in vitro assays can be used to evaluate the ADMET characteristics of potential leads at earlier stages of development. This is done in order to eliminate candidates as early as possible, thus avoiding costs, which would have been expended on chemical synthesis and biological testing. Scientists are developing computational methods to select only compounds with reasonable ADMET properties for screening. Molecules from these computationally screened virtual libraries can then be synthesized for high-throughput biological activity screening. As the predictive ability of ADME/Tox software improves, and as pharmaceutical companies incorporate computational prediction methods into their R&D programs, the drug discovery process will move from a screeningbased to a knowledge-based paradigm. 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.

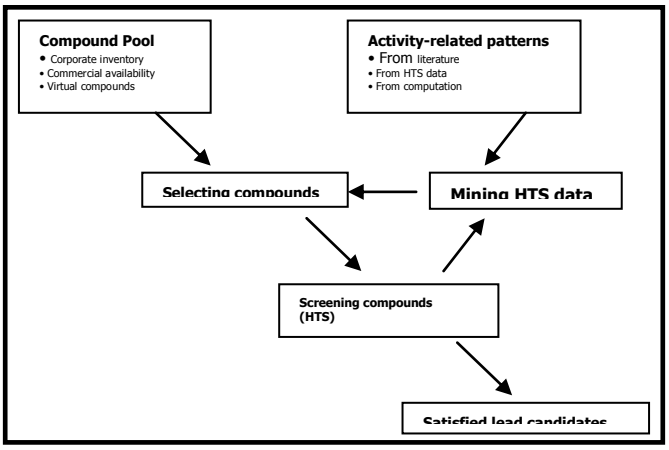

Fig. 2 Sequential High Throughput Screening

Futuristic use of chemo-informatics and CAMD

In the past decade this two techniques have developed vastly and attained many achievements in the field of drug discovery. The newer

challenge for this techniques is the multi-parametric optimization technique for the *in silico* optimization of ADMET parameters of the lead compounds along with their HTS via predictive computational models. The molecules from these computationally screened virtual libraries can then be synthesized for high-throughput biological activity screening. Many new technologies such as support vector machines (SVM) have found recent scientific applications. SVM may be able to eliminate many problems that are encountered through other approaches such as decision trees or neural networks.

References

Augen, J. *Drug Discov. Today*, **2002**, 7, 315.

Baxter, A.D.; lockey, P.M. *Drug Discov. World*, **2001**, 2, 9.

Bernard, P.; Golbraikh, A.; Kireev, D.; Chretien, J.R.; Rozhkova, N. *Analysis*, **1998**, 26, 333.

Brown, R.D.; Martin, Y.C. *J. Chem. Inf. Comput. Sci.*, **1996**, 36, 572.

Cox, T.F.; Cox, M.A.A. *Multidimensional* ***Scaling,*** Chapman & Hall/CRC Press: Boca Raton, **2000**.

Cramer, R.D., III; Patterson, D.E.; Bunce, J.D. *J. Am. Chem. Soc.*, **1988**, 110 (18), 5959.

Galop, M.A.; Barrett, R.W.; Dower, W.J.; Fodor, S.P.A.; Gordon, E.M. *J. Med. Chem.*, **1994**, 37, 1233.

Hecht, P. *Curr. Drug. Discov.*, **2002**, 21.

http://esc.syrres.com/interkow/doesmile.htm

http://statsoft.com/textbook/stfacan.html

http://www.asis.org/Features/Pioneers/wiswess.htm.

http://www.statsoft.com/textbook/stmulsca.html#general

ISIS/Base help file "Remote QB in a molecule Database: Searching Concepts/Examples".

Jarvis, R.A.; Patrick, E.A. *IEEE T. Comput.*, **1973**, C22, 1025.

Joliffe, I.T. *Principal Component* ***Analysis***, Springer-Verlag: New York, **1986**.

Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeny, P.J. *Adv. Drug Deliv. Rev.*, **1997**, 23, 3.

Malinowski, E.H.; Howery, D.G. *Factor Analysis in* ***Chemistry***, John Wiley & Sons: New York, **1980**.

Martin, Y.C. In *Methods in* ***Enzymology***; Lilley, D.M.J.; Dahlberg, J.E., Eds.; Academic Press: San Diego, **1991**; pp. 587-613.

Matter, H.; aringhans, K.H.; Naumann, T.; Klabunde, T.; Pirard, B. *Comb. Chem. High T. Sci.*, **2001**, 4, 453.

Organic 'Tectons' used to make networks with Inorganic Properties. *Chem. Eng. News*, **1995**, 21.

Rebek, J., Jr. *Acc. Of Chem. Res.*, **1990**, 23 (12), 39.

Sadowski, J.; Kuninyi, H.A. *J. Med. Chem.*, **1998**, 41, 3325.

Walters, W.p.; Murcko, M.A. *J. Med. Chem.*, **1998**, 41, 3314.

Weiner, H. *J. Am. Chem. Soc.*, **1947**, 69, 17.

Weininger, D. *J. Chem. Inf. Comput. Sci.*, **1988**, 28, 31.

Willett, P.; *Similarity and Clustering in Chemical Information* ***Systems,*** Research Studies Press, Wiley: New York, 1987.

Xu, J.; Stevenson, J. *J. Chem. Inf. Comput. Sci.*, **2000**, 40, 1147.

Xue, L.; Stahura, F.L.; Godden, J.W.; Bajorath, J. *J. Chem. Inf. Comput, Sci.*, **2001**, 41, 394.

Zupan, J.; Gasteiger, J. *Neural Networks for* ***Chemists***, VCH: Weinheim, 1993.

Xu, J.; Hagler, A. *Molecules*, **2002**, 7, 566.