Cheminformatics Approaches in Modern Drug Discovery

9

Salma Jamal and Abhinav Grover

9.1 Traditional Drug Discovery and Bottlenecks

Drug discovery is the process by which novel candidate chemical compounds are identified and validated as medications for the treatment of specific disorders and diseases. Traditionally, drugs were discovered through identification of active components from traditional therapies without prior knowledge of the biological target. This was later replaced by chemical repositories of natural compounds or synthetic small molecule compounds which were screened against cell lines or organisms for the identification of compounds having desirable therapeutic properties. The traditional drug discovery process includes various steps: selection of disease, identification of target and its validation, searching lead compound effective against target and its synthesis, preclinical testing in animal models and human clinical trials [1] as is evident from Fig. 9.1. However, these steps are associated with serious bottlenecks which include the large amount of time and costs involved in making and testing of the novel candidate entities identified. This led to the need of newer technologies that could automate the drug discovery process such as high throughput screening (HTS) where millions of chemical compounds could be screened per drug target per year [2]. In order to meet the increased requirement of novel chemical compounds, combinatorial chemistry technologies, which made it possible to make millions of chemical compounds in one go, were started to be used [3]. Conversely, it resulted in disappointing results as this process could not yield significant drug

S. Jamal
School of Biotechnology, Jawaharlal Nehru University, New Delhi 110067, India

Department of Bioscience and Biotechnology, Banasthali University,
Tonk, Rajasthan 304022, India

A. Grover (✉)
School of Biotechnology, Jawaharlal Nehru University, New Delhi 110067, India
e-mail: abhinavgr@gmail.com

A. Grover (ed.), *Drug Design: Principles and Applications*,
DOI 10.1007/978-981-10-5187-6_9

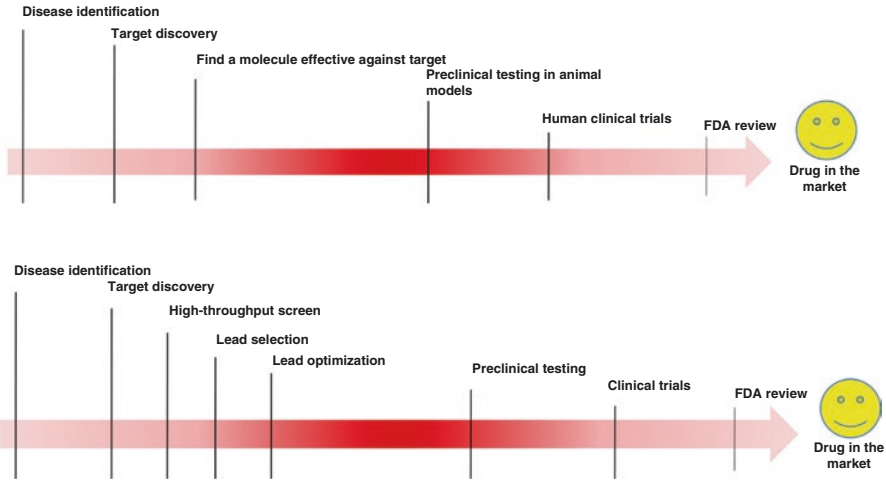

Fig. 9.1 Shows the traditional drug discovery process and modern process of drug discovery and development

candidates due to lack of chemical diversity as well as drug-like properties leading to wastage of large number of compounds. Another problem to be addressed was the identification of candidate molecules with desirable therapeutic properties from the large pool of compounds. In order to overcome these drawbacks, a method was needed which could predict drug-like compounds as well as absorption, distribution, metabolism, excretion and toxicity (ADMET) properties of the screening compounds and thus various cheminformatics, sometimes referred to as chemoinformatics or chemical informatics, approaches were introduced. Cheminformatics approaches are widely used by scientists, researchers and pharmaceutical as well as other chemical and related industries in the drug discovery process in addition to a plethora of other applications which would be discussed further.

9.2 An Introduction to Cheminformatics

F.K. Brown defines chemoinformatics as

"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" [4] as is shown in Fig. 9.2.

As quoted by W. Warr at http://www.warr.com/warrzone2000.html,

"Chem(o)informatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information" by G. Paris (August 1999 Meeting of the American Chemical Society) [5].

Various other definitions have been given for cheminformatics that include "Chemoinformatics – A new name for an old problem" by M. Hann and R. Green

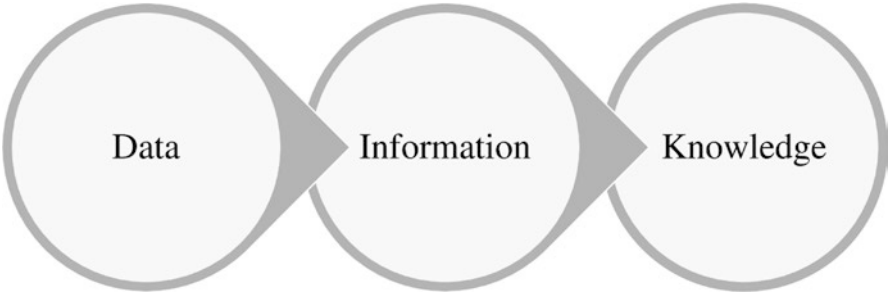

Fig. 9.2 Shows the flow of information in cheminformatics

[6] and "The application of informatics methods to solve chemical problems" by J. Gasteiger and T. Engel [7]. Cheminformatics can be broadly defined as use of information technologies and computers to solve chemical problems such as data mining, information retrieval and extraction, topology of chemical compounds and various others. Cheminformatics has been known to mainly deal with small molecules which involves analysis of scientific data and extraction of information to assist in development of new compounds, where bioinformatics, in addition to large chemical compounds, deals with genes and proteins. However, both the approaches, cheminformatics and bioinformatics, are adjunct to each other in providing us deeper insights for biomolecular processes such as prediction of structure and function of proteins, ligand binding to active sites and enzyme catalysis. A perfect example of this can be seen in drug design process where bioinformatics methods are used for identification of targets for novel drug candidates and cheminformatics methods are used for finding these new small molecule drug candidates [5].

9.3 Cheminformatics, Its Importance and Various Aspects

The major role of cheminformatics is to store, search and handle enormous amounts of chemical data which comprises ever increasing number of millions of chemical compounds generated every year. Furthermore, it also includes extraction of information and knowledge from this chemical data which could be used to model the relationships between chemical structures and biological activities and predict the bioactivities of other chemical compounds from their structures. The four important problems solved by cheminformatics include storing a molecule, searching the exact molecule in a database, substructure searching and similarity search. Based on these, cheminformatics has three major aspects which are: information acquisition, information management and information use (Fig. 9.3).

Information acquisition deals with the generation and collection of experimental and theoretical data. Recent years have seen advancements in high throughput screen technologies and combinatorial synthesis which has enabled generation and analysis of large number of chemical compounds which could be tens or hundreds or thousands or even millions of molecules in a very short time period. Since such a

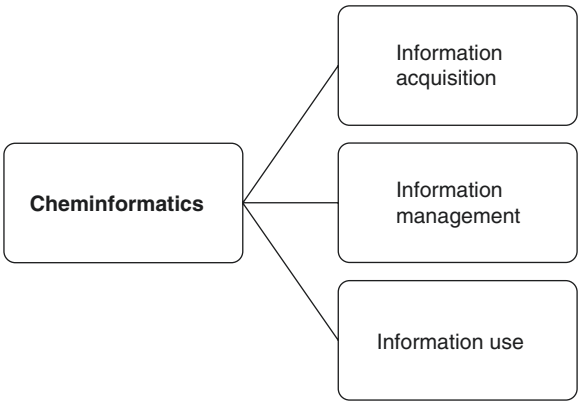

Fig. 9.3 Shows the various aspects of cheminformatics

huge amount of information is being generated, a novel technique is needed which could store and analyse this information for an effective purpose and the answer is Cheminformatics [8].

Information management deals with the storage and retrieval of the chemical information. Various methods which could result in two- and three-dimensional representation of conformations of chemical structures and similarity searching have been discovered. Also, this information is now been incorporated into chemical information storage systems using which molecules with high percentage of features matching to target the molecule can be searched and retrieved [8].

Information use includes analysis of data as well as its application to various biochemical problems. Using the chemical informatics tools and techniques, the raw data collected can now be successfully accessed and used to obtain valuable results such as prediction of unknown properties of chemical compounds consequently leading to development of novel candidate drug-like compounds [8].

9.4 Cheminformatics Approaches

Various approaches have been followed for the effective implementation of cheminformatics which include chemical structure representation, selection of descriptors, modelling of properties, classification and pattern recognition, virtual screening and structure- and ligand-based drug design.

9.5 Chemical Structure Representation Via Descriptors or Features

One major challenge for cheminformatics is the correlation of chemical information with biological data for which appropriate description of chemical structures is very important. Descriptors or features are the properties of an object which can be measured and if that object is a molecule, descriptors may be defined as the

mathematical representation of the chemical information encoded within the molecule. Todeschini and Consonni defined molecular descriptors as, "The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment" [9]. Various features of the chemical compounds can be quantified which include theoretical properties such as number of various atoms such as hydrogen, oxygen, etc., donors and acceptors, number of rotatable bonds, etc., as well as experimental properties like log *P*, polarizability, and many more. Molecular descriptors can be computed using a number of commercial and free molecular descriptor generation software which include ADAPT [10], DRAGON [Talete, Milano, Italy], ADMET Predictor [Simulations Plus Inc., Lancaster, CA], JOELib (JOELib/JOELib2 cheminformatics library), Molecular Operating Environment (MOE, Chemical Computing Group) (Chemical Computing Group Inc. 2015), MARVIN beans [ChemAxon], PowerMV [11], PaDEL [12] and many more.

9.6 Descriptor Selection for Dimension Reduction

Not all the descriptors chosen to represent the molecule are relevant for predicting the biological activity of the compounds which on the other hand increase the dimension of the matrix, say for a library of *n* compounds, *m* descriptors are chosen where *m* could be any number, the resulting matrix would be a large *n***m* dimensional matrix [1]. Also the accuracy and robustness of the models generated to predict the bioactivity of compounds depend on the descriptors chosen to represent the molecules [13]. Thus in order to reduce the dimensions and noise as well as to identify significant descriptors, descriptor selection is needed. Another problem is the repetition of the descriptors for the entire library of the compounds, say a feature *m* has same values for all the *n* compounds of a library. Thus this feature is not providing any significant information for the prediction task and could be removed without having any impact on the accuracy of prediction [14]. Descriptor or feature selection is a technique to find an ideal subset of features, from the set of original features, which may be less in number but are not redundant, exceedingly important and have most contribution towards the prediction [15]. The remaining features are considered irrelevant and are discarded. An optimum subset of features comprises descriptors which are related to the bioactivity of the compounds, are highly informative, are independent of each other, and are easy to understand and noise free [16].

9.7 Classification Using Machine Learning Methods

Machine learning methods are very popular in cheminformatics and have been widely used for classification of active compounds and prediction of unknown properties of compounds. Machine learning is based on learning from a set of training

instances with known properties and then predicting values for unknown properties of test instances. Tom M. Mitchell's defined machine learning as "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [17]. As quoted from Arthur Samuel's definition of machine learning in 1959, machine learning is a "field of study that gives computers the ability to learn without being explicitly programmed" [18]. Various methods/algorithms have been known to be used for the learning purposes such as artificial neural networks, decision-tree based learning, Bayesian models, support vector machines and k-nearest neighbours and many more. We have discussed the aforementioned machine learning algorithms in the following section.

9.8 Artificial Neural Network

Artificial neural networks (ANN), also known as neural networks, are a computational approach which mimic the biological neural network of the brain and work the way human brain solves the problems with large clusters of neurons connected with axons. The term "network" in ANN refers to the interconnected neurons in different layers of the system. The system works using three layers, input layer which is the first layer consisting of input neurons and sends data, via synapses¸ to the second layer which is a hidden layer and this hidden layer further sends the data to the third and the last output layer. The number of layers may increase with the complexity of the systems. The synapses which connect these layers are associated with weights that are manipulated during the learning phase. Basically, three parameters are used to define ANN which are: the pattern that interconnects the different layers, the weights connected to these layers which are fluctuated during learning and the activation function responsible for the conversion of weighted input neurons to outputs [19]. ANN have been widely used for a range of cheminformatics applications such as in the work on steroids by So et al. [20], study of oestrogen receptor agonists by Li et al. [21], and also by Briem et al. [22] to identify possible kinase inhibitors and many more to count.

9.9 Random Forest

Random forest (RF) is a decision-tree based technique which uses an ensemble of trees for the classification task. The training data consisting of multiple features for each instance is used to build the decision trees. A multitude of decision trees are constructed during the training phase and the output class is the mode of the classes output of the individual trees. The trees consist of nodes, branches and edges which correspond to the features, values and classes, respectively. One feature is selected randomly at each node which separates the objects/instances to classes with maximum information gain. There is no pruning of the trees and each tree is grown to the largest possible extent. The tree is terminated when each of the features has

been considered at least once or if the feature gives same value for all the training objects. The trained forest of trees is then used for the prediction of future unseen data [23]. RF method has been successfully used in cheminformatics for investigation into mutagenicity data [22], development of hERG blocker classifiers [24], quantitative-structure activity relationships models (QSAR) [25] and skin sensitization data [26].

9.10 Naive Bayes

The naïve bayes (NB) classifier is a probabilistic classifier which uses Bayes theorem and classifies the test instances to the class with highest probability. The algorithm assumes that all the features are conditionally independent of each other and thus have impartial contribution towards the task of prediction irrespective of any correlation among the features [27]. The advantage of NB classifiers over other classification algorithms is its simplicity and that it's extremely fast. A NB model is easy to generate and does not involve any complex parameter tuning which makes it highly efficient for handling large datasets. NB classifiers have performed quite well and are frequently used in cheminformatics, for target prediction [28], classification of drug-like molecules based on their biological activities [29], for toxicity prediction [30] and various other studies.

9.11 Support Vector Machines

Support vector machine (SVM) is a non-probabilistic binary linear classifier that separates the instances belonging to two classes using a gap which is as wide as possible. This gap is defined by a separating hyperplane which is the output of the SVM algorithm and categorizes new instances to the class depending on which side of the hyperplane these instances fall. SVM can also efficiently perform non-linear classification using kernel functions. Some common kernel functions include polynomial, Gaussian radial basis function, and hyperbolic tangent [31]. SVMs are one of most widely used methods in cheminformatics and have been used for prediction of toxicity [30], classification of kinase inhibitors [22], mutagenic toxicity prediction [32] and prediction of biological activities of small molecules and drug-repurposing [33].

9.12 K-Nearest Neighbours

The k-nearest neighbour algorithm (kNN) is based on the principle that the categorization of an instance to a particular class depends on the majority of votes of its neighbours. For example, to predict the class of an instance, x, the classifier will look for k-nearest neighbours of instance x and assign it the class to which most of the k-nearest neighbours belong. K-nearest neighbours are selected on

the basis of the distance between the test instance and training instances in the feature space. Euclidean distance is usually used for this purpose; however, other metrics like Jaccard distance could also be used [34]. Many studies have used kNN classification algorithm for the prediction of anticancer drugs [35], psychoactivity of cannabinoid compounds [36], and mutagenicity of chemicals [37] and several others.

9.13 Applications of Cheminformatics

There is a wide range of applications of cheminformatics starting from storage and retrieval of chemical compounds to prediction of their properties, QSAR, virtual screening and drug design and various others like textile industry, molecular, material and food science and combinatorial organic synthesis [38]. We have discussed some of the typical applications of cheminformatics.

9.14 Storage and Retrieval of Chemical Compounds

This is the primary application of cheminformatics which involves storing chemical compounds information generated through experiments and retrieval of information and structures from chemical databases [39]. Table 9.1 provides the list of the databases for the storage of information on chemical compounds.

Table 9.1 Provides the list of the databases for the storage of information on chemical compounds

PRIVATE Database	Supplier
ACD (Available Chemicals Directory)	MDL Information Systems Inc.
ASINEX	AsInEx Ltd.
CDD (Collaborative Drug Discovery)	Collaborative Drug Discovery, Inc.
ChEMBL	European Bioinformatics Institute (EBI)
ChemIDplus	U.S. National Library of Medicine, National Institutes of Health
ChemSpider	Royal Society of Chemistry
CSD (Cambridge Structural Database)	Cambridge Crystallographic Data Centre
InterBioScreen	InterBioscreen Ltd.
Maybridge	Thermo Fisher Scientific Inc.
MedChem	Daylight Chemical Information Systems Inc.
NCI96 (National Cancer Institute data)	Daylight Chemical Information Systems Inc.
PubChem	National Centre of Biotechnology Information
Spresi95	InfoChem GmbH
TSCA93	Daylight Chemical Information Systems Inc.
WDI (World Drug Index)	Derwent Publication
ZINC database	University of California San Francisco

9.15 Prediction of Properties of Chemical Compounds

One of the most important tasks for chemists is to generate compounds with desirable properties. A large number of compounds generated through HTS and CC technologies had to be discarded since these compounds did not have drug-like properties. This wastage could be minimized if we had a technique to predict physical, chemical or biological properties of the compounds and synthesize compounds with properties which could result in novel drug-like candidates [5].

9.16 QSAR

QSAR method involves the prediction of properties of chemical compounds from their structures. QSAR models comprise predictors which consist of physiochemical properties or theoretical molecular descriptors of chemical compounds and predict the biological activities of the chemicals. The models first establish a relationship between structures of chemical compounds and biological activities and based on that predict the activities of novel chemical compounds [40].

9.17 Virtual Screening

Virtual screening has become one of the important tools for identifying drug-leads. The technique involves computational screening of large in silico libraries of compounds to identify compounds with desirable properties such as having activity against a biological target and filtering unwanted compounds [41]. Various virtual screening methods are applied that include docking if the target structure is known [42], similarity approaches if the target structure is unknown but the ligands are known [43] and structure–activity relationship approaches if neither the structure of the target nor the ligands structure is known [44].

9.18 Cheminformatics and Modern Drug Discovery

Recent years have seen large amount of chemical data generated through integration of CC and HTS technologies for the purpose of drug discovery. However, this data can be effectively utilized only if we have techniques to store, handle, analyse and apply it in the drug discovery process. The traditional drug discovery process involved identification of disease and drug target followed by synthesis of molecule effective against that disease. This molecule would further be investigated for pharmacodynamics and pharmacokinetic properties and toxicity and then taken up for clinical trials. This is a costly and lengthy process with the risk of the lead molecules being failed in the clinical trials leading to wastage of time, money and efforts. This highlights the need for the technique which could identify the problematic lead compounds and predict the biological activities and ADMET properties of the

chemical compounds before the preclinical testing thus reducing the rate of failures and speeding up the process of drug discovery and development. Cheminformatics is one such technique which plays a major role in identification of drug target and lead compounds active against the drug target and prediction of their ADMET properties. The modern drug discovery process involves four steps: identification of target and its validation, lead identification, lead optimization followed by preclinical trials (Fig. 9.1).

Once the target is identified, the lead compounds with desirable properties active against that target could be screened out from the enormous number of diverse chemical compounds available through cell-based assay compounds or various databases of small molecules owing to the HTS technique. Various cheminformatics approaches like machine learning could be used to generate computational models which could identify novel drug candidates from lead compounds. Further, the selected candidate compounds could be docked to the protein target to find out the compounds having affinity towards the target. When the drug-like compounds have been identified, these could be taken forward to evaluate ADMET characteristics using computational models thus eliminating the undesirable compounds at earlier stages of drug development and cutting down the costs and time involved. Various other cheminformatics techniques like similarity searching and substructure searching could be applied for the identification of novel scaffolds from large chemical compounds repositories.

9.19 Tools and Techniques Used in Cheminformatics

Various cheminformatics toolkits that can be used by cheminformaticians for chemical data mining, virtual screening and structure–activity relationship studies have been developed. We have listed some of the tools below:

- ChemDraw [45] is a Macintosh and Microsoft windows program first developed by David A. Evans and Stewart Rubenstein in 1985 and later by the cheminformatics company CambridgeSoft. This is a molecular editor tool which, along with Chem3D and ChemFinder, is part of the ChemOffice suite of programs.
- ChemReader is a fully automated tool that extracts chemical structures information from images in research articles and translates that information into standard chemical formats that can be searched and analysed [46].
- ChemSketch is a molecular modelling program that allows drawing and modification of structures of chemical compounds and structural analysis that includes understanding of chemical bonds and functional groups [47].
- ChemWindow is a program developed by Bio-Rad Laboratories, Inc. that allows drawing chemical structures, 3D visualization and database searching.
- Chemistry Development Kit (CDK) is a JAVA software for use in bioinformatics and cheminformatics available for Windows, Linux and Macintosh. The program allows 2D molecular generation, 3D geometry generation, descriptors and fingerprints calculation and supports various chemical structure formats [48].

- ChemmineR is a R language cheminformatics program for analysing small molecule drug-like compounds data and enables similarity searching, clustering and classification of chemical compounds using a wide range of algorithms [49].
- JME molecular editor is a JAVA applet that allows to create and modify chemical compounds and reactions and can display molecules within an HTML page [50].
- Molecular Operating Environment is a scientific vector language based software program the applications of which include structure- and fragment-based design, pharmacophore synthesis, protein and molecular modelling and simulations in addition to cheminformatics and QSAR.
- Open Babel is a software which is used for the interconversion of chemical file formats. It also allows substructure searching as well as fingerprints calculation [51]. It is available for Windows, Linux and Macintosh.
- OpenEye is a drug discovery and design software kit and its areas of application include generation of chemical structures, docking, shape comparison, cheminformatics and visualization. OpenEye toolkits are available in multiple programming languages that are C++, JAVA and python.
- Chemaxon provides various cheminformatics software programs, applications and services for drawing structures of chemical compounds and their visualization, searching and management of chemical databases, clustering of chemical compounds and drug discovery and design.
- PubChem is a large repository of chemical compounds and their biological activities obtained through biological assays. The database is maintained by National Centre for Biotechnology Information and is freely accessible [52].

Various other tools such as PowerMV, PaDEL, CDD (Collaborative Drug Discovery), RDKit, 3D-e-chem, ADMET Predictor, MedChem Studio, MedChem Designer, Mol2Mol, Chimera, VMD, ArgusLab, ChemTK, Premier Biosoft and many others are also widely used for cheminformatics applications. Table 9.2 lists the companies that provide cheminformatics software and tools.

Table 9.2 Lists the companies that provide cheminformatics software and tools

Accelrys, USA	CambridgeSoft, USA
Advanced Chemistry Development, Canada	Bioreason, USA
Agilent Technologies, USA	Aventis Pharma (France, Germany, USA)
Bayer (Germany, USA)	Novartis Pharma, USA
Chemical Diversity Labs, USA	Molsoft, USA
Daylight USA	ACD/Labs (Advanced Chemistry Development, Inc.)
Golden Helix Inc., USA	Schrödinger, USA
ID Business Solutions Ltd., United Kingdom	OpenEye Scientific Software, USA
Modgraph Consultants Ltd., United Kingdom	Molinspiration, Slovak Republic
Molecular Networks, Germany	Eidogen-Sertanty, USA
PharmaDM, Belgium	SciTegic, USA
Rosetta Biosoftware	MolMine Bioinformatics Software Solutions

Conclusion

The large amount of HTS data generated in recent years has triggered the need for developed cheminformatics systems. Cheminformatics methods enable us to manage and understand increased amount of chemical data, and analyse and exploit the results obtained from experiments. In addition to chemical information, cheminformatics methods also allow to retrieve information about physical properties, spectroscopic data, crystallographic and 3D molecular structures, chemical reaction pathways, functional groups, docking and various other parameters. Cheminformatics has made the process of drug discovery and development very easy and fast. In this chapter, we have discussed about cheminformatics and its various aspects and approaches, traditional drug discovery process and role of cheminformatics in modern drug discovery system. Cheminformatics has already been successfully integrated into various fields such as chemistry, bioinformatics and many more; however, there still is a need for developed and advanced cheminformatics systems for solutions to unsolved problems.

Acknowledgements AG is thankful to Jawaharlal Nehru University for usage of all computational facilities. AG is grateful to University Grants Commission, India for the Faculty Recharge Position. Salma Jamal acknowledges a Senior Research Fellowship from Indian Council of Medical Research (ICMR), New Delhi.

Competing Interests The authors declare that they have no competing interests.

References

1. Xu J, Hagler A (2002) Chemoinformatics and drug discovery. Molecules 7:566–600
2. Hecht P (2002) High-throughput screening: beating the odds with informatics-driven chemistry. Curr Drug Discov:21–24
3. Gallop MA, Barrett RW, Dower WJ, Fodor SP, Gordon EM (1994) Applications of combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries. J Med Chem 37:1233–1251
4. Brown FK (1998) Chemoinformatics: what is it and how does it impact drug discovery. Annu Rep Med Chem 33:375–384
5. Engel T (2006) Basic overview of chemoinformatics. J Chem Inf Model 46:2267–2277
6. Hann M, Green R (1999) Chemoinformatics—a new name for an old problem? Curr Opin Chem Biol 3:379–383
7. Gasteiger J, Engel T (2006) Chemoinformatics: a textbook. Wiley
8. James CA Cheminformatics 101. An introduction to the computer science and chemistry of chemical information systems. eMolecules Inc., Del Mar
9. Todeschini R, Consonni V (2008) Handbook of molecular descriptors, vol 11. Wiley, NewYork
10. Valla A, Giraud M, Dore JC (1993) Descriptive modeling of the chemical structure-biological activity relations of a group of malonic polyethylenic acids as shown by different pharmacotoxicologic tests. Pharmazie 48:295–301
11. Liu K, Feng J, Young SS (2005) Power MV: a software environment for molecular viewing, descriptor generation, data analysis and hit evaluation. J Chem Inf Model 45:515–522
12. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474

13. Mitchell JB (2014) Machine learning methods in chemoinformatics. Wiley Interdiscip Rev Comput Mol Sci 4:468–481
14. Alpaydin E (2014) Introduction to machine learning. MIT Press, Cambridge
15. Daumé H (2012) A course in machine learning (ciml.Info), p. 189
16. Brown RD, Martin YC (1996) Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. J Chem Inf Comput Sci 36:572–584
17. Mitchell TM (1997) Machine learning. McGraw-Hill Science/Engineering/Math, Maidenhead, p. 432
18. Simon P (2013) Too big to ignore: the business case for big data. Wiley, Hoboken, p. 89
19. Mitchell JBO (2014) Machine learning methods in chemoinformatics. Wiley Interdiscip Rev Comput Mol Sci 4:468–481
20. So S-S, Karplus M (1997) Three-dimensional quantitative structure− activity relationships from molecular similarity matrices and genetic neural networks. 1. Method and validations. J Med Chem 40:4347–4359
21. Li H et al (2006) Prediction of estrogen receptor agonists and characterization of associated molecular descriptors by statistical learning methods. J Mol Graph Model 25:313–323
22. Briem H, Günther J (2005) Classifying "kinase inhibitor-likeness" by using machine-learning methods. Chembiochem 6:558–566
23. Jehad Ali RK, Ahmad N, Maqsood I (2012) Random forests and decision trees. Int J Comput Sci Issues 9
24. Marchese Robinson RL, Glen RC, Mitchell JB (2011) Development and comparison of hERG blocker classifiers: assessment on different datasets yields markedly different results. Mol Informat 30:443–458
25. Kuz'min VE, Polishchuk PG, Artemenko AG, Andronati SA (2011) Interpretation of QSAR models based on random forest methods. Mol Informat 30:593–603
26. Li S, Fedorowicz A, Singh H, Soderholm SC (2005) Application of the random forest method in studies of local lymph node assay based skin sensitization data. J Chem Inf Model 45:952–964
27. Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. Mach Learn 29:131–163
28. Koutsoukas A et al (2013) In silico target predictions: defining a benchmarking data set and comparison of performance of the multiclass Naïve Bayes and Parzen-Rosenblatt window. J Chem Inf Model 53:1957–1966
29. Cannon EO et al (2007) Support vector inductive logic programming outperforms the naive Bayes classifier and inductive logic programming for the classification of bioactive chemical compounds. J Comput Aided Mol Des 21:269–280
30. von Korff M, Sander T (2006) Toxicity-indicating structural patterns. J Chem Inf Model 46:536–544
31. Platt JCSequential minimal optimization. A fast algorithm for training support vector machines. Report no. MSR-TR-98-14, 21 (Microsoft Research), 1998)
32. Liao Q, Yao J, Yuan S (2007) Prediction of mutagenic toxicity by combination of recursive partitioning and support vector machines. Mol Divers 11:59–72
33. Kinnings SL et al (2011) A machine learning-based method to improve docking scoring functions and its application to drug repurposing. J Chem Inf Model 51:408–419
34. Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 46:175–185
35. Ajmani S, Jadhav K, Kulkarni SA (2006) Three-dimensional QSAR using the k-nearest neighbor method and its interpretation. J Chem Inf Model 46:24–31
36. Honório KM, da Silva AB (2005) A study on the influence of molecular properties in the psychoactivity of cannabinoid compounds. J Mol Model 11:200–209
37. Basak SC, Grunwald GD (1995) Predicting mutagenicity of chemicals using topological and quantum chemical parameters: a similarity based study. Chemosphere 31:2529–2546

38. Begam BF, Kumar JS (2012) A study on cheminformatics and its applications on modern drug discovery. Proced Eng 38:1264–1275
39. Aktar MW, Murmu S (2008) Chemoinformatics: principles and applications. 1 Pesticide Residue Laboratory, Department of Agricultural Chemicals, 2 Department of Agricultural Chemistry and Soil Science, Bidhan Chandra Krishi Viswavidyalaya, Mohanpur-741252, Nadia, West Bengal, India.
40. Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V (2009) A practical overview of quantitative structure-activity relationship. EXCLI J 8:74–88
41. Walters WP, Stahl MT, Murcko MA (1998) Virtual screening—an overview. Drug Discov Today 3:160–178
42. Diller DJ, Merz KM (2001) High throughput docking for library design and library prioritization. Proteins 43:113–124
43. Willett P (2000) Chemoinformatics–similarity and diversity in chemical libraries. Curr Opin Biotechnol 11:85–88
44. Gedeck P, Willett P (2001) Visual and computational analysis of structure–activity relationships in high-throughput screening data. Curr Opin Chem Biol 5:389–395
45. Halford B (2014) Reflections on CHEMDRAW. Chem Eng News 92:26–27
46. Park J et al (2009) Automated extraction of chemical structure information from digital raster images. Chem Cent J 3:4
47. Hunter AD (1997) ACD/ChemSketch 1.0 (freeware); ACD/ChemSketch 2.0 and its Tautomers, Dictionary, and 3D Plug-ins; ACD/HNMR 2.0; ACD/CNMR 2.0. ACS Publications.
48. Steinbeck C et al (2003) The chemistry development kit (CDK): an open-source Java library for chemo-and bioinformatics. J Chem Inf Comput Sci 43:493–500
49. Cao Y, Charisi A, Cheng L-C, Jiang T, Girke T (2008) Chemmine R: a compound mining framework for R. Bioinformatics 24:1733–1734
50. Ertl P (2010) Molecular structure input on the web. J Cheminform 2(1)
51. O'Boyle NM et al (2011) Open babel: an open chemical toolbox. J Chem 3:33
52. Wang Y et al (2009) PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 37:W623–W633