Data mining and fusion methods in ligand-based virtual screening

Mubarak Himmat[1], Naomie Salim[1], Mohammed Mumtaz Al-Dabbagh[1], Faisal Saeed and Ali Ahmed[1,2]
[1]Faculty of Computing, Universiti Teknologi Malaysia, Skudai, Johor, 81310, Malaysia.
[2]Faculty of Engineering, Karary University, Khartoum 12304, Sudan.
***Corresponding author: E-Mail: barakamub@yahoo.com**

ABSTRACT

Computational methods in drug discovery have increasingly gained attention of researchers within the last decades, as the tools for facilitating drug discovery process, for it allow rapid screening of huge databases , the ligand-abased virtual screening (LBVS) methods continue to be developed and improved as one of the important tool of drug discovery process ,and it is become one of the most interested area in Chemo informatics, this review discuss various methods that are applied for LBVS ,the paper focus deeply in the various of machine learning techniques that have been applied for LBVS , and also it will demonstrate and discuss the effectiveness of data fusion in LBVS ,and how these techniques provided effectiveness tools could enhance the LBVS .

KEY WORDS: Virtual screening, drug discovery, Machine learning, data mining.

1. INTRODUCTION

Data mining methods and techniques have been applied and proposed in many aspects of the sciences, as it provides sophisticated solutions that have increased the influence of this information technology in real life within different fields. Data mining methods and information retrieval methods have been applied in chemical, biomedical and other medical fields. Text mining is used to extract information automatically form sources (written documents) by using computational methods (Jensen, 2006); this extraction will generate new information, and has also been widely used and applied in the identification of some disease-associated entities like genes and proteins, which could help with understanding their roles in diseases. One of the articles that proposed text mining in this domain is the work done by (Ozgur, 2008), where proposed and automatic literature mining-based methods included text mining approaches that predict good candidate genes before conducting real experiments. On the other hand, there are many methods that facilitate these techniques by combining several biological concepts with computers and new IT techniques and tools, and statistical methods that can be used to discover and extract beneficial information.

In drug discovery, which is considered one of the most complex and costliest processes, there are considerable efforts and powerful techniques that have been made to develop and simplify the process of drug discovery. The actual laboratory drug discovery process can take between 12 and 15 years and can cost approximately one million dollars (Rollinger, 2008); for that, considerable effort has been made to cover research into this area. This has taken years and cost in excess of $1 billion; it is complex and costly and consumes a lot of time in laboratory experiments. Nowadays, the data mining some machine learning approaches techniques have become some of the most important basic steps in the process of drug discovery and have led to significant advances in this area and it is play a major role in Chemo informatics.

During the process of drug discovery, a lot of work and insensitive investigations are performed to ensure the identification of similar molecules or biological therapeutics; these are defined as development candidates, which, if successful, will contribute to clinical development and be marketed as medicines. Within this chapter, we will focus on using machine learning and fusion techniques in LBVS and discuss many approaches and proposed methods.

2. VIRTUAL SCREENING

The process of discovering of new drugs using computational screening methods is being continuously devolved and improved as it is one of the most important tools for drug discovery and an alternative to high-throughput biochemical compound screening (HTS). HTS was considered the basic and a main method for drug candidate development, but LBVS and its various techniques and search methods is becoming a reliable method for drug discovery.

The basic idea of drug discovery is to identify new chemical entities that have the ability to bind to a target protein and extract the desired biological response. In the last few decades, with the use of computer technology and methods facilitating numerous aspects of research in all fields, computer applications have become some of the most important tools to assist chemists. They are reliable methods and techniques and can be used in many aspects of chemistry, such as molecule ranking, clustering, docking and virtual screening; as a result, this is now used as a complementary tool to HTS in drug discovery. VS methods are proposed to speed up the process of drug discovery and to increase the efficiency and reduce the high costs. Here, a huge amount of databases can be screened easily and successfully in a short time.

VS or screening, as described here, describes the process of selecting molecules to help in bioactivity testing. This screening is applied automatically by computer methods that select molecules; this is generally referred to as VS.

Most of the molecules in VS do not actually exist in molecule stores; the process screens a virtual library of molecular stores. The screening methods conducted by computers are employed to rank the molecules according to their structure and put the most promising structures at the top of the list; this gives a high-ranking to those molecules with structures that may be similar to structures that have already been tested. The screening methods and concept of molecular similarity are closely related to those used in information retrieval. VS is now becoming more accepted as an effective method in lead discovery, since it provides a good method that can be used to eliminate undesired molecules from compound libraries, which directly influence the cost of drug discovery and contribute to the reduction of time and cost of molecules and the recovery of new drug discovery projects. The application of screening methods has become a basic tool for drug discovery, and is considered a technique that classifies and identifies whether the molecule tends to be active or inactive in a biochemical test.

Similar property principle: When concentrating on LBVS and ranking methods in Chemo informatics, the basic principle is that molecules based on two compounds that are 'similar' to one another in their structure will have similar properties and activities; this principle is helpful for the identification of similar molecules in databases of molecules descriptors.

Chemical database searching: VS techniques in drug discovery rely on searching databases of molecules to find similar molecules to the query; search methods in VS are classified into three classes or types: Structure searching, Substructure searching and Similarity searching. Most researches have proposed methods that could provide better results in this searching and also in the ranking of molecules and clustering.

The wider area of Chemo informatics and especially the similarity searching is concerned with enhancing searching methods and the ways to calculate similarity; these similarity calculations and methods not only used in LBVS, but they are also used in other Chemo informatics aspects, like property prediction, structure based searching, molecular diversity analysis and synthesis design. The calculation process of similarity depends on many factors that must be taken into account when talking LBVS, including the molecular representation and similarity coefficients that are used to compare molecules, and so forth.

In structure searching, the search is conducted to find out the specific structure of the molecule in the database; the query must find an exact match of the structure query, which is called a structure search (Girschick, 2013; Lyne, 2002; Waszkowycz, 2008; Barnard, 1993). The second type of research is substructure searching, where the search looks at a database to identify the structure or molecules that contain one or more particular structural fragments of the query. Much work has been done in substructure searching, and still continues; the work suggested by (Barnard, 1993; Willett, 2009; Bender, 2009) has applied different methods of substructure searching. The third search type is similarity searching, where the search will look for all structures in a database that achieve a high similar to a given structure. For this type of search, there are many different similarity measures and similarity coefficients which have been applied and used (Girschick, 2013; Downs, 1994; Downs and Willett, 1996; Willett, 2003; 2006; Salim, 2003). Most of these similarity measures have been derived from techniques that were already used in the area of text information retrieval. Also, there are a variety of similar similarity searching methods which have been proposed (Bender, 2009; Holliday, 2011; Jahn, 2009). In chemical data similarity, searching is usually done using fingerprint representations with different types of coefficients.

Ligand based virtual screening: There is no doubt that LBVS has become an important source and reliable tool for the drug discovery process and plays an essential role in increasing and enhancing the drug discovery process. The search for active compounds using computational methods is becoming sophisticated, especially in the area of Chemo informatics. LBVS relies on using known and active compounds as input information and then trying to find any of the structurally diverse compounds that have similar bioactivity. There are a lot of methods that have been proposed and applied in ligand-based VS. A wide range of research in VS has focused on enhancing, evaluating and comparing the proposed methods, and the result of the recall of known active compounds the resulted by screening methods is considered the major measure of successful method performance. Until now, VS has lacked standards for method evaluation in general (Geppert, 2010). Also it does not agree one hundred percent with the real live performance in practical applications; the identification of some active compounds that are structurally similar and separate from the available reference molecules is considered a successful screen. Although this area is still rich, there are many problems that need to be solved and a lot of methods need to be enhanced to achieve good performance in all VS approaches.

Recently, many data mining machine learning methods have been applied in both LBVS and structure based screening, but these methods are not applied for predictive models, their concepts are just adapted for use; especially for VS purposes, many machine learning techniques and data mining methodologies have been increasingly applied to identify active compounds and have become viable alternatives to compound classification. Similarity search methods and conventional structure-activity relationship analysis use machine learning methods such as Support Vector Machine (SVM), k-Nearest Neighbours, Random Forest and Naïve Bayes; also, there are other techniques for data mining which are applied in different area of Chemo informatics, and most of them have been applied in LBVS, such as Bayesian methods, Decision trees and Voting techniques.

Screening Process on Databases: Ligand-based virtual screening relies on knowledge of the active query molecules; to conduct this virtual screening, we must prepare the query and databases of molecules and searching methods. This is considered a basic step of virtual screening, as chemical compound databases must be prepared before starting to use search methods. The preparation is done according to the requirements of the selected descriptors: 2D fingerprint or 3D descriptors, and then employing the proper virtual screening methods. There are many searching methods applied in similarity searching, meaning that there is no unique method or way to quantify the similarity between molecules (queries and references); however, the most common way is to use similarity coefficients, and there are a lot of similarity measures applied in virtual screening which have been discussed in the literature (Todeschini, 2012; Piwowarski, 2010). Also, there are some other methods like machine learning methods (Hert, 2006; Chen, 2009) and fusion methods (Whittle, 2006; Willett, 2006; 2013; Ahmed, 2014).

The screening is conducted after selecting one of the proposed similarity methods; the active query ligand is used to search databases with a specific descriptor to look for molecules in databases that are most like the query. The result of each search is a list of compounds with similarity values; this list is then ranked according to the similarity of the reference compound values in decreasing order, and then the user sees how many of the active compounds have been recalled in the database that contains some active and non-active molecules. The advantage of LBVS is that the results of searches will provide small subsets of compounds that can be quantified in real throughput screening.

Data mining machine learning techniques in LBVS: In LBVS, there are various methods that have been applied, as mentioned before, to improve the search efficiency of screening, these methods work with special search algorithms and different molecule descriptors. Machine learning approaches enter the field of virtual screening. Here we will discuss some common methods that have been applied in LBVS. The simplest case is a similarity search where a single active query molecule and an enumerated search database are required. More complex are similarity searches in combinatorial spaces where the enumeration of all search database molecules is not possible due to their sheer number. This case requires special search algorithms and molecule descriptors. Finally, for machine learning approaches, the knowledge of many active and inactive molecules is required in order to train models to categorize the actives and the inactive molecules in the search database. also the machine learning has been used as data fusion , for it used to define the optimum combination of similarity coefficient that could be used for data fusion for specific class of active compounds (Chen, 2009).

Machine-learning methods in VS: This section discuss the most common LBVS methods, and approaches ,firstly we will start by machine learning approaches ,the machine learning has been applied in different aspects of computer sciences ,and it aimed to design programs that has ability to learn and/or discover, and it could automatically improve the performance on certain tasks, in Chemo informatics machine learning methods has been enter for two purposes ,for compounds classification, and for ligand based-virtual screening ,in virtual screening it has been used for analysing the structural characteristics of molecules of known active and inactive molecule, where the availability of these known active and inactive sets that used as training sets and used as tool that could apply to the unknown molecules the test set to predict their the active and inactive molecules and enhancing the screening results. Several machine learning (Hert, 2006; Chen, 2009; Grant, 2006; Jorissen and Gilson, 2005; Han, 2008) and methods have been developed, also the machine learning has been used as data fusion, for it used to define the optimum combination of similarity coefficient that could be used for data fusion for specific class of active compounds (Chen, 2009).

Support Victor Machine: Recently, support Vector Machines (SVM) approaches have been interested machine learning methods that attracted a lot of attention of researchers for it provide acceptable and high prediction accuracy in different aspects, the support victor machine(SVM) is one of the common machine learning approaches In machine learning, SVM is supervised learning models with associated learning algorithms that used for analyse data and recognize patterns, and it is applied as classification and regression analysis. It has been used by giving a sets of training that each of these set is marked for belonging to one of two categories, SVM algorithms has been applied to solve virtual screening problem and to enhance the screening recall.

And it have been used as LBVS tools that facilitated the process of lead discovery ,one of the early work of using SVM is the work that done by (Jorissen and Gilson, 2005) they used SVM as tools that solve problem of enriching a database of molecules for active molecules. Their proposed model of SVM model generates substantial enrichment of active molecules with chemistries different from those in the training set, they depend rely on molecular descriptors of the active and inactive compounds in a training data set and used it as train a Support Vector Machine, and the descriptors of the molecules used to be points in a multidimensional space where each dimension corresponds to one of the descriptors. The SVM machine method here responsible to find a boundary that best separates the two sets of points corresponding to the active and inactive compounds. Then the final result of their SVM model ranks a test set that consists of other active and inactive, the performance of the model is measured by the recall of the active compounds that have been resulted. Another work done by (Han, 2008) proposed and SVM screening method for large compound libraries, their proposed methods produce lower hit-rate compared by other methods of VS tools at that time and it recorded best performing, their method partly because their training-sets contain limited spectrum of inactive compounds.

Bayesian methods: As we mentioned before there are many different methods an approaches that have been applied in LBVS and fingerprint-based molecular similarities searching, and most of these method are inspired for the research that have been done in textual information retrieval area, The Bayesian inference networks have been used in deferent areas; it enables prediction of an event accruing according to some probability computations, allowing for the fact that this chosen event can be dependent on other events occurring. The Bayesian inference networks has been applied as similarity searching tool of chemical compounds tools, and it have been applied in and evaluated in LBVS by (Chen, 2009), after they did some modifications to the Bayesian techniques that used in information retrieval ,for they used The Bayesian inference networks to rank the molecular database in decreasing order of probability of bioactivity, many works adapted the Bayesian methods to be used in LBVS ,and to be used as alternative similarity based virtual screening (Abdo, 2010; Bender, 2011), the work that done by (Abdo, 2010) demonstrated clearly the Bayesian inference networks provided good result in LBVS especially with structural homogenous molecules.

Decision trees: A decision tree is one of the machine learning techniques that used as is a decision support tool, and it has been applied in different aspect of the sciences, it represents and uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, costs, resource, and utility. It is one way to display an algorithm in Chemo informatics decision tree has been used in classification rather than virtual screening, and rare works have been adapted for virtual screening using (Plewczynski, 2006).

3. DATA FUSION

Data fusion in virtual screening is the process of combining different screening results with several reference ligands and/or several search methods (Salim, 2003; Hert, 2006; Ginn, 2002), and there are two type of fusion that applied in virtual screening, similarity fusion (Chen, 2009) and group fusion (Whittle, 2006), the similarity fusion is the combination of similarity searches for that used one reference ligand with different descriptors as similarity fusion, and the group fusion is the combination of the results that obtained for a set of reference ligands with one method. Recently many ligands based virtual screening new methods and approaches have been applied. In early works of data fusion in virtual screening they just proposed fusion method of combination similarity coefficients (Salim, 2003; Ginn, 2002; Holliday, 2002), their proposed methods suggested fusion of similarity measures by combination of the screening results that achieved by using multiple similarity measures (Chen, 2009). The most of these ideas of fusion are derived from Information retrieval of combining of ranking procedures that applied in textual data (Belkin, 1995), and they applied the same techniques in their work they use MAX, MIN rules. The ranking based fusion which is proposed by (Ginn, 2002) is considered as appropriateness of rank-based fusion. Their rankings are fused using the SUM, MIN and MAX. Another work that done by (Willett, 2006) applied anew approach of fusion where they machine-learning in similarity searching as fusion method, they prepare a "training set" of known active and non-active molecules, and use the developed machine learning tools to predict the active molecules form the unknown activity (the test set), also they proposed a new fusion by using machine learning methods in combination with references structure. There are some other fusion types that have been described in (Whittle, 2006). They used different fusion types like similarity fusion by using different Coefficients and also similarity fusion by using different rules proposed by (Chen, 2010). In their work they found that an analysis of their fusion rule is effectiveness. The studies by (Whittle, 2004) showed that the group fusion results are better than the traditional similarity searching and the best results were found by using the MAX fusion rule. All the works mentioned achieved good results, but we must keep in mind that no fusion of the coefficients and rules can be expected to get better results all the time. Recently a new approach has been proposed, this new approach is combining the both ligand-based and structure-based in virtual screening (Drwal and Griffith, 2013), other new Anew approach method of fusion are proposed by (Ahmed, 2014), they proposed Condorcet fusion ,the fusion is conducted by combines the outputs of similarity searches using several association and distance similarity coefficients, and then try to find the best measure based on Condorcet fusion to be the winner measure for each class of molecules, a lot of new works (Willett, 2013; Cano, 2014) are predict the use of fusion in the future.

4. CONCLUSION

Drug discovery has many challenges, but there have been great interest from researchers to solve these challenges by including many disciplines in life sciences and informatics. LBVS has presented many solutions and is continuously trying to provide helpful tools and methods for drug discovery. This paper has given an overview of drug discovery basic concepts; and we basically focused on ligand-based LBVS and how data mining methods have helped in drug discovery. Then, we discussed the different chemical molecule searching methods. In particular, we highlighted the most well-known of the LBVS approaches, and the fact that have been observed is researches in this area are increasing, and in the near future, it is hypothesised that there will be considerable changes in drug discovery using VS methods and techniques, because drug design and development is costly and takes a long time; therefore, computational methods and VS methods will become widely used and reliable approaches in drug discovery.

5. ACKNOWLEDGEMENTS

This work is supported by the Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under the Research University Grant Category (VOT Q.J130000.2728.01K73).

REFERENCES

Abdo A, Ligand-based virtual screening using Bayesian networks, Journal of chemical information and modeling, 50(6), 2010, 1012-1020.

Ahmed A, Condorcet and borda count fusion method for ligand-based virtual screening, Journal of Chemo informatics, 6(1), 2014, 1-10.

Barnard J.M, Substructure searching methods: old and new, Journal of Chemical Information and Computer Sciences, 33(4), 1993, 532-538.

Belkin N.J, Combining the evidence of multiple query representations for information retrieval, Information Processing & Management, 31(3), 1995, 431-448.

Bender A, Bayesian methods in virtual screening and chemical biology, in Chemo informatics and Computational Chemical Biology, Springer, 2011, 175-196.

Bender A, How similar are similarity searching methods? A principal component analysis of molecular descriptor space, Journal of chemical information and modeling, 49(1), 2009, 108-119.

Cano G, García-Rodríguez J, and Perez-Sanchez H, Improvement of Virtual Screening Predictions using Computational Intelligence Methods, Letters in Drug Design & Discovery, 11(1), 2014, 33-39.

Chen B, Mueller C, and Willett P, Combination Rules for Group Fusion in Similarity-Based Virtual Screening, Molecular Informatics, 29(6-7), 2010, 533-541.

Chen B, Mueller C, and Willett P, Evaluation of a Bayesian inference network for ligand-based virtual screening, Journal of chemo informatics, 1(1), 2009, 1-10.

Chen J, Holliday J, and Bradshaw J, A machine learning approach to weighting schemes in the data fusion of similarity coefficients, Journal of chemical information and modeling, 49(2), 2009, 185-194.

Downs G.M, and Willett P, Similarity searching in databases of chemical structures, Reviews in computational chemistry, 7, 1996, 1-66.

Downs G.M, Willett P, and Fisanick W, Similarity searching and clustering of chemical-structure databases using molecular property data, Journal of Chemical Information and Computer Sciences, 34(5), 1994, 1094-1102.

Drwal M.N, and Griffith R, Combination of ligand-and structure-based methods in virtual screening, Drug Discovery Today Technologies, 10(3), 2013, e395-e401.

Geppert H, Vogt M, and Bajorath Jr., Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation, Journal of chemical information and modeling, 50(2), 2010, 205-216.

Ginn C.M, Willett P, and Bradshaw J, Combination of molecular similarity measures using data fusion, in Virtual Screening: An Alternative or Complement to High Throughput Screening, Springer, 2002, 1-16.

Girschick T, Puchbauer L, and Kramer S, Improving structural similarity based virtual screening using background knowledge, Journal of chemo informatics, 5(1), 2013.

Grant J.A, Lingos, finite state machines, and fast similarity searching, Journal of chemical information and modeling, 46(5), 2006, 1912-1918.

Han L, A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor, Journal of Molecular Graphics and Modelling, 26(8), 2008, 1276-1286.

Hert J, New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching, Journal of chemical information and modeling, 46(2), 2006, 462-470.

Holliday J.D, Hu C, and Willett P, Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. Combinatorial chemistry & high throughput screening, 5(2), 2002, 155-166.

Holliday J.D, Multiple search methods for similarity-based virtual screening: analysis of search overlap and precision, Journal of chemo informatics, 3(1), 2011, 1-15.

Jahn A, Optimal assignment methods for ligand-based virtual screening, Journal of chemo informatics, 1, 2009, 14.

Jensen L.J, Saric J, and Bork P, Literature mining for the biologist: from information retrieval to biological discovery, Nature reviews genetics, 7(2), 2006, 119-129.

Jorissen R.N, and Gilson M.K, Virtual screening of molecular databases using a support vector machine, Journal of chemical information and modeling, 45(3), 2005, 549-561.

Lyne P.D, Structure-based virtual screening: an overview, Drug Discovery Today, 7(20), 2002, 1047-1055.

Ozgur A, Identifying gene-disease associations using centrality on a literature mined gene-interaction network, Bioinformatics, 24(13), 2008, i277 - i285.

Piwowarski B, What can quantum theory bring to information retrieval, in Proceedings of the 19th ACM international conference on Information and knowledge management, ACM, 2010.

Plewczynski D, Spieser S.A, and Koch U, Assessing different classification methods for virtual screening, Journal of chemical information and modeling, 46(3), 2006, 1098-1106.

Rollinger J.M, Stuppner H, and Langer T, Virtual screening for the discovery of bioactive natural products, in Natural Compounds as Drugs, Volume I, Springer, 2008, 211-249.

Salim N, Holliday J, and Willett P, Combination of fingerprint-based similarity coefficients using data fusion, Journal of chemical information and computer sciences, 43(2), 2003, 435-442.

Todeschini R, Similarity coefficients for binary chemo informatics data: overview and extended comparison using simulated and real data sets, Journal of chemical information and modeling, 52(11), 2012, 2884-2901.

Waszkowycz B, Towards improving compound selection in structure-based virtual screening, Drug discovery today, 13(5), 2008, 219-226.

Whittle M, Analysis of data fusion methods in virtual screening: similarity and group fusion, Journal of chemical information and modeling, 46(6), 2006, 2193 – 2205, 2206-2219.

Whittle M, Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: a comparison of similarity coefficients, Journal of chemical information and computer sciences, 44(5), 2004, 1840-1848.

Willett P, Combination of similarity rankings using data fusion, Journal of chemical information and modeling, 53(1), 2013, 1-10.

Willett P, Enhancing the Effectiveness of Ligand-Based Virtual Screening Using Data Fusion, QSAR & Combinatorial Science, 25(12), 2006, 1143-1152.

Willett P, Fusing similarity rankings in ligand-based virtual screening, Computational and Structural Biotechnology Journal, 2013, 5.

Willett P, Similarity methods in chemo informatics, Annual review of information science and technology, 43(1), 2009, 1-117.

Willett P, Similarity-based approaches to virtual screening, Biochemical Society Transactions, 31(3), 2003, 603-606.

Willett P, Similarity-based virtual screening using 2D finger prints, Drug discovery today, 11(23), 2006, 1046-1053.