VOL. 10, NO. 3, FEBRUARY 2015 ISSN 1819-6608

ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

DATA FUSION APPROACHES IN LIGAND-BASED VIRTUAL SCREENING: RECENT DEVELOPMENTS OVERVIEW

Mubarak Himmat[1], Naomie Salim[1], Ali Ahmed[1, 2], and Mohammed Mumtaz Al-Dabbagh[1]
[1]Faculty of Computing, University of Technology Malaysia, Skudai, Johor, Malaysia
[2]Faculty of Engineering, Karary University, Khartoum, Sudan
E-mail: barakamub@yahoo.com

ABSTRACT

Virtual screening has been widely used in drug discovery, and it has become one of the most wealthy and active topic areas in Chemoinformatics. Virtual screening (VS) plays a major role in drug discovery process, for the process of drug discovery is costly, Virtual screening has been used to reduce this cost, recently, there are many different virtual screening methods that have been suggested and applied on chemical databases. This paper aims to discuss theoretically the VS approaches, and searching methods, and demonstrates the recent approaches of VS. It's meanly focus and discuss the issue of using data fusion and how it increases the screening performance level, and demonstrate the different types of fusions that are applied in VS, discussing and exploring the enhancements and effectiveness that happen with applying the different types of applied fusion techniques, and discuss future trends of virtual screening.

Keywords: Chemoinformatics, data fusion, similarity searching, virtual screening, drug discovery**.**

INTRODUCTION

Information retrieval and data screening now become a wide and wealth area of computer science and it is research areas. Several searching mechanisms are developing every day, in different areas. In Chemoinformatics, information retrieval techniques are also applied since last four decades, the enhancing of retrieving data is becoming one of most concern of the researchers,

A wide variety of methods and algorithms have been suggested and developments in virtual screening to capture the chemical similarity between compounds, the screening are lead and contributed in the process of drug discovery [1-3] .The Ligand-based virtual screening is now representing a powerful tool for exploring and identifying new biologically active compounds. In ligand-based virtual screen, the search is done by one or more reference active molecule, these active molecules are comparing the database, the recall of similar molecules in the database will be top-ranking and returned then evaluated, and tested for biological activity.

To enhance virtual screening(VS) new approach have been suggested by using Data fusion ,The data fusion in virtual screening are first time proposed by Chemoinformatics group of University of Sheffield [4-6], ,and in clustering also some fusion are applied [7], and now it is going increasingly. As VS is one of the major aspects of Chemoinformatics and drug discovery, and there are many methods and algorithms that have been applied to search in chemical compound data sets for and find the active molecules, to help laboratories detect active molecules. In recent decades and since the earliest of the 2000th the data fusion is applied in the Chemoinformatics, and some of these fusions are suggested and applied in the visual screening and clustering of chemical databases. Each process of VS is focus on searching for molecules in the database in order to calculating a similarity score for each compound structure in the molecule databases and then rank these molecules with their structures in decreasing order of the already calculated scores, to achieve the similarity principle that indicates the same molecule structure, are expected to behave the same biological activity of reference structure.

This survey presents the latest methods applied in virtual screening, and it generally focuses on proposed methods of data fusion, those methods and techniques used to enhance the virtual screening. The rest of this paper is organized as follows: In the next section we will introduce a review of the previous and later works that have been undertaken in virtual screening, and it is common approaches, the next section will discuss the chemical compound database search and it categorises, and then we will discuss the data fusion in virtual screening, then we will go into a deep discussion of data fusion methods that have been applied in virtual screening, and finally the paper is concluded.

VIRTUAL SCREENING APPROACHES

Searching and screening similarities among objects is now concerned in several areas such as information retrieval, word computing, fault diagnosis, clustering analysis, ranking and detecting duplicate documents, plagiarism detection, comparing user profiles in a social media application, searching for similar topics in literature, and so forth. In Chemoinformatics searching for molecule and determining the similarity among them .Now , VS has become a common approach in the process of discovering new drugs, the efficiency of this drug discovery depends on the screening tools' performance and the database that has been used in the screening, generally there was two approaches in virtual screening as we mentioned , Ligand-based and structured –based,

Similarity searching techniques, ligand-based VS on a chemical database has become widely used, and VS plays an essential role and dependant method of drug discovery now VS has new approaches that uses combination of Ligand-based and structured based method Figure-1 below illustrate the three VS approaches . Much work has been done in similarity searching. In Chemoinformatics, the VS is built on the principle of similarity, which means the process of both active subset selection and structure activity relationship studies postulate that compounds similar to biologically active ones should also be active and vice versa. And also the similarity principle relies on the fact that proteins that share similar sequences are considered to have similar structures and exhibit the same or similar functions.

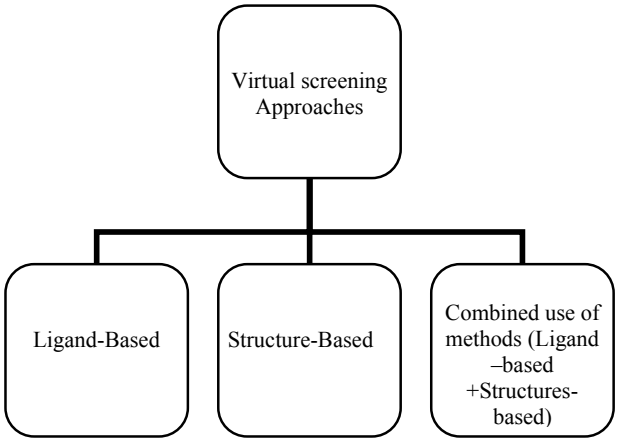

Figure-1. Virtual screening approaches.

CHEMICAL DATABASE SEARCH

This section will introduces VS, similarity search and data fusion and points to related works. So far, many works have been done in the VS of chemical databases. The VS has become a common approach in the process of discovering new drugs, molecules and compounds. The efficiency of this drug discovery depends on the screening tools' performance and the database that has been used in the screening**.**

Similarity searching techniques for ligand-based VS on a chemical database has become widely used, and VS has become an essential and significant element in the drug discovery process. Much work has been done in similarity searching. In Chemoinformatics, the VS is built on the principle of similarity, which means the process of both active subset selection and structure activity relationship studies postulate that compounds similar to biologically active ones should also be active and vice versa. And also the similarity principle relies on the fact that proteins that share similar sequences are considered to have similar structures and exhibit the same or similar functions.

Any large molecule consists of small molecules that are called ligands. The VS is target has been used for the purpose of finding the novel ligand that will match the human body proteins and be bind to. Generally, the process of VS starts from the molecules that have the same known activities and use it to identify similar molecules. It can be divided into two classes: ligand-based and target-based VS In this paper, we will focus on using ligand-based VS, more than focus on target-based VS.

The cause of seeking for the similar molecules is that for Molecular Similarity Principle that has states that most structurally similar molecules are tend to have similar properties of both physicochemical as well as biological ones , The chemical molecule screening process are conducted by applying similarity coefficients, these similarity coefficients are applying to the molecules database to work as functions that calculate the similarity between two molecules, and then transform their representations into real numbers, a lot of similarity measures are applied in binary system and in Chemoinformatics, one of the recently works that done by [8]compare and discussed about 50^{th} different similarity measures that have been applied on chemical molecule dataset.,a different ways are proposed to enhancing the screening result ,using similarity coefficients and others assistance of decryption tools , [9] propose a similarity searching technique depend on the environment of the atoms, their technique is by Appling naive Bayesian classifier and feature selection in different datasets and then they compare their works with alternative searching methods.

There are three different type of searching, In this section we will discuss the three ways of searching in a chemical compounds.

Structure searching

This type of search is looking for a specific structure of the molecule in the database, and the result of this query will be finding the exact match of molecular structure in databases. There are two types of molecules, represented by 2D and 3D. Diagram (1) shows the structure of the toms present in the alcohol molecule in 2D representation, and the bond connections between them. Suppose we were searching about the following query "alcohol", the result of the search will only found and retrieve an exact match to this structure (alcohol), The process of the search will try to match the query structure of (alcohol) in the 2D connection table as a mathematical graph, where each node represents atoms and the edges represent bonds, and the process testing the matching query will be done by using the a graph algorithm and techniques that are provided in information retrieval and computer science, recent work done by [10] proposed improvement of structure based similarity searching by unextending the background knowledge of binding of compounds ant the it result significant improvement in both similarity measure and rankings virtual.

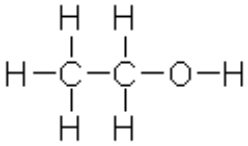

Figure-2. representation of alcohol structure.

Substructure searching

Substructure searching search is the search that executes to find out the structures or the molecules that contains one or more particular structural fragments of query. For example, we have "alcohol" molecules and we want to find all the compounds that have the structures of "alcohol" in a database or to find the "alcohol" group the substructure searching will return all compounds contain alcohol as the part of its components. The result of the search can find many compounds, many works are covered Substructure searching [11, 12].

Similarity searching

Similarity searching in one of the important search method In Chemoinformatics many different similarity measures have been applied and suggested the article discussed most coefficients that applied in similarity measure of molecule searching [8], in 1990th Willet[1] provide and review on the using the similarity coefficients in search in chemical database, the idea of using these similarity is derived from techniques that were already used in the area of the binary system and textual information retrieval. The screening here will conduct to find a similar molecules by look for all the structures in a database that are achieving the highly similar to a given structure[13]. The reason for using the similarity search is to find compounds that could exhibit similar properties, as in Chemoinformatics the basic idea of the "similar property principle" that fond the compounds with similar structures are likely to exhibit similar biological behaviours'. The similarity search are applied in in 2D fingerprint datasets [14],and in the other types of dataset, similarity search are used for clustering of data [15, 16] .As an example, a similarity search might involve looking for structures with a similarity greater than the fixed percentage the specific molecule. Similarity searching is usually done using fingerprint representations and similarity different types of coefficients.

DATA FUSION

The term fusion means "the integration of information from multiple source to produce a specific and comprehensive unified data about entity" [17]. The first time that data fusion applications were used was in the 1980s by the United States Department of Defence. The fusion system was used to track and find military targets like tanks. The system collects data from several sensors and then combines the information for accuracy purposes and to increase efficiency, but currently the usage of data and information fusion is becoming more popular and it has been applied to a large number of fields, for improvements that it provided, and the common sense lead to get that multi-sources are better than one resource, and as people say "one hand can't clamp", and this principle has led to activating the use of data fusion. In Chemoinformatics the data fusion (or the consensus scoring) means is the process of combine the results of different similarity searches of a chemical database compound ,at end of 1990th the work of [18] use data fusion of combine ranking of several different measures of inter molecule structure ,that work is considered one of the firsts of using fusion in virtual screening, and also this approach is called similarity fusion, for similarity fusion in virtual screening execute the searching by a single reference structure but at the same time using multiple similarity measures as the work that done by [5, 6] They combine several binary similarity coefficients and then took the overall searching result of the similarity, most of the data fusion in virtual screening is relay on the idea of computing a score (approximating, directly or indirectly, to the probability of activity), this computing are conduct for each molecule in a database by using multiple scoring functions. Then these multiple sets of scores are combined to obtain a better set of scores than that which obtained by the using of single function[19].For instance, we can conduct a combine searches for a specific reference structure by that had been executed with more than one different types of fingerprints or of similarity measure, The other type of fusion is Group fusion, were as we mentioned before the search execute by searching with similarity measure but using multiple reference structures[20].Data fusion is based on the manipulation of multiple measurements and it is implemented by applying some arithmetic operations on predefined lists of similarity scores (or ranks) which result from the number of searches. These arithmetic operations could be coefficients or rules like (MAX, MIN, SUM,…).The fused score for MIN , MAX,SUM could be expressed as follow :

$$\text{MIN}\{SIM1(dy), SIM2(dy)..SIMx(dy)..SIMn(dy)\},$$

$$\text{MAX}\{SIM1(dy), SIM2(dy)..SIMx(dy)..SIMn(dy)\},$$

$$\sum_{x=1}^{n} SIM_x(d_y)$$

Min,Max rules are assign the y-th database-structure, dy, a score that is the minimum (in Min) or large (in Man)of the n similarities to the reference structure that calculated by the above rules;

SUM rule assigns by a score that is the sum of the n individual similarities. Many rules are discussed in the article [21]. In Chemoinformatics, data fusion is applied to structure-based and ligand-base, but in this article we focus on fusion in VS which has attracted the attention of a lot of researchers. As general data fusion has two approaches, *similarity fusion* and *group fusion*, in similarity fusion the search will be provided by the type of research that start researching by single reference structure with multiple similarity measures. This approach has been applied in deferent aspects of computer science [7]. The group fusion is vice versa of similarity fusion. In group fusion the searching starts with searches with a single similarity measure and using multiple reference structures ,many research proposed different group fusion ,by applying different rules [19, 21-23] . The group fusion results found that it is more effective than similarity fusion [21]. Group fusion is applied by using fusion rules. There are many data fusion approaches which have been developed for chemical data and for ligand-based VS.

DISCUSSIONS

With the growing of importance of the using computer techniques in drug discovery, researchers started mining to present useful and reliable methods that result strong contributions in the drug discovery area. Scholars and researchers try to apply some methods that have been used in binary systems, text, images, and other area in Chemoinformatics, these research founding led of using data fusion in Chemoinformatics and in it is both aspect clustering and Virtual screening, It is now become considerable that data fusion is present an effective result in virtual screening, with valuable result that obtained by most of the proposed work in virtual

At the beginning of this article we discussed and investigated the current study of data fusion, and we found that, so far, a large number of individual VS methods have been examined and tested. Most of them used different data fusion. Data fusion has now become one of the most widely used methods for enhancement of VS, and there are many data fusion approaches that have been developed for ligand-based VS in Chemoinformatics started in early 1990s, and then increased throughout the subsequent years. This became clear when we look at the recent articles in Chemoinformatics; we found that there are a lot of data fusion articles, most of recent article's authors recommend using data fusion. And recently also there are many works that have been done using fusion techniques. In early works of data fusion articles they just proposed fusion method of combination similarity coefficients [5, 6, 24], they suggest similarity measures by combining the results of searching based on multiple similarity measures, they also propose methods used for the making combination of many individual search methods and provided multi-similarly measures, Other combination of the result that done by searching using different similarity coefficients are proposed by [25] but the combination in their work was quite different from other combination of similarity coefficients for the depend on the size, in terms of sub-structural fragments present. All these idea and methods are derived from Information retrieval of combining of ranking procedures that applied in textual data [26], and they applied the same techniques in their work they use MAX, MIN rules. The ranking based fusion which is proposed by [6] is considered as appropriateness of rank-based fusion. Their rankings are fused using the SUM, MIN and MAX. Another work [19] applied a new group fusion rule to enhance the similarity searching. They use machine learning tools with data fusion. In their work the machine-learning methods firstly need to prepare a "training set". The training set of known active and non-active molecules, and use the developed machine learning tools to apply the molecules of unknown activity (the test set) to predict there in actives. Their study used different references and also they described a new fusion by using machine learning methods in combination with references structure. There are some other fusion types that have been described in [22]. They used different fusion types like similarity fusion by using different Coefficients and also similarity fusion by using different rules proposed by [21]. In their work they found that an analysis of their fusion rule is effectiveness. The studies by [4] 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 [27], other new Anew approach method of fusion are proposed by[28] ,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 [29, 30] are predict the use of fusion in the future.

CONCLUSIONS

As we mentioned there are many different similarity measures are proposed for ligand based virtual screening and virtual screening as all, using these coefficients as single provided a good result, but using data fusion either similarity fusion or group fusion give better results, and increase the efferent of screening results.

Most of new articles in screening are concern of using data fusion in VS is a way of enhancing the similarity search and ranking of molecules. In this article we introduced the concepts of using data fusion in VS and similarity search. The results of the research indicate that combination ligand-based and structure-based approach and data fusion have a good contribution in improving retrieval performances. Moreover, the data fusion applied in Chemoinformatics different aspects similarity measures,

clustering and VS will not be restricted to only Chemoinformatics. It is embedded and involved in complex data mining, clustering and so forth and all the works those have been applied and presented are good, but it can be considered by the limitation and the accuracy of the VS scoring functions. As we have seen, fusion of data could provide more significant advantages than using single source data, and also the results of applying data fusion increase the efficiency and confidence of a similarity search and also in classification results. This means that the use of fusion will be the more concerned by researchers in the near future, and this statement has been recommended by many scholars in the field of Chemoinformatics. Recently many articles have focused on data fusion. The current importance of data fusion indicates a resurgence of interest in the use and further development of data fusion. We therefore conclude that data fusion could be extended, and more fusion methods could be achieved in future. In summary, The data fusion in Virtual screening is still rich area that needs more digging and mining, and it represents big challenges, more work and further developments are expected in future.

ACKNOWLEDGEMENTS

This work is supported by the Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under Research University Grant Category (VOT Q.J130000.2528.07H89).

REFERENCES

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

[2] D. Horvath, "A virtual screening approach applied to the search for trypanothione reductase inhibitors," Journal of medicinal chemistry, vol. 40, pp. 2412-2423, 1997.

[3] P. Willett, J. M. Barnard. and G. M. Downs, "Chemical similarity searching," Journal of chemical information and computer sciences, vol. 38, pp. 983-996, 1998.

[4] M. Whittle, V. J. Gillet, P. Willett, A. Alex, and J. Loesel, "Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: a comparison of similarity coefficients," Journal of chemical information and computer sciences, vol. 44, pp. 1840-1848, 2004.

[5] N. Salim, J. Holliday, and P. Willett, "Combination of fingerprint-based similarity coefficients using data fusion," Journal of chemical information and computer sciences, vol. 43, pp. 435-442, 2003.

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

[7] T. Lange and J. M. Buhmann, "Fusion of similarity data in clustering," in NIPS, 2005.

[8] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, and P. Willett, "Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets," Journal of chemical information and modeling, vol. 52, pp. 2884-2901, 2012.

[9] A. Bender, H. Y. Mussa, R. C. Glen, and S. Reiling, "Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance," Journal of chemical information and computer sciences, vol. 44, pp. 1708-1718, 2004.

[10] T. Girschick, L. Puchbauer, and S. Kramer, "Improving structural similarity based virtual screening using background knowledge," Journal of cheminformatics, vol. 5, p. 50, 2013.

[11] G. B. McGaughey, R. P. Sheridan, C. I. Bayly, J. C. Culberson, C. Kreatsoulas, S. Lindsley, et al., "Comparison of topological, shape, and docking methods in virtual screening," Journal of chemical information and modeling, vol. 47, pp. 1504-1519, 2007.

[12] J. M. Barnard, "Substructure searching methods: old and new," Journal of Chemical Information and Computer Sciences, vol. 33, pp. 532-538, 1993.

[13] P. Willett, "Similarity-based approaches to virtual screening," Biochemical Society Transactions, vol. 31, pp. 603-606, 2003.

[14] P. Willett, "Similarity-based virtual screening using 2D fingerprints," Drug discovery today, vol. 11, pp. 1046-1053, 2006.

[15] P. S. Charifson, J. J. Corkery, M. A. Murcko, and W. P. Walters, "Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins," Journal of medicinal chemistry, vol. 42, pp. 5100-5109, 1999.

[16] G. M. Downs, P. Willett, and W. Fisanick, "Similarity searching and clustering of chemical-structure databases using molecular property data," Journal of Chemical Information and Computer Sciences, vol. 34, pp. 1094-1102, 1994.

[17] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," Proceedings of the IEEE, vol. 85, pp. 6-23, 1997.

[18] J. Bradshaw, "The Application Of Data Fusion To Similarity Searching In Chemical Databases," 1998.

[19] P. Willett, "Enhancing the Effectiveness of Ligand Based Virtual Screening Using Data Fusion," QSAR & Combinatorial Science, vol. 25, pp. 1143-1152, 2006.

[20] A. Bender and R. C. Glen, "Molecular similarity: a key technique in molecular informatics," Organic & biomolecular chemistry, vol. 2, pp. 3204-3218, 2004.

[21] B. Chen, C. Mueller, and P. Willett, "Combination Rules for Group Fusion in Similarity Based Virtual Screening," Molecular Informatics, vol. 29, pp. 533-541, 2010.

[22] M. Whittle, V. J. Gillet, P. Willett, and J. Loesel, "Analysis of data fusion methods in virtual screening: similarity and group fusion," Journal of chemical information and modeling, vol. 46, pp. 2206-2219, 2006.

[23] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, et al., "Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information," Journal of medicinal chemistry, vol. 48, pp. 7049-7054, 2005.

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

[25] J. Chen, J. Holliday, and J. Bradshaw, "A machine learning approach to weighting schemes in the data fusion of similarity coefficients," Journal of chemical information and modeling, vol. 49, pp. 185-194, 2009.

[26] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw, "Combining the evidence of multiple query representations for information retrieval," Information Processing & Management, vol. 31, pp. 431-448, 1995.

[27] M. N. Drwal and R. Griffith, "Combination of ligand-and structure-based methods in virtual screening," Drug Discovery Today: Technologies, vol. 10, pp. e395-e401, 2013.

[28] A. Ahmed, F. Saeed, N. Salim, and A. Abdo, "Condorcet and borda count fusion method for ligand-based virtual screening," Journal of Cheminformatics, vol. 6, pp. 1-10, 2014.

[29] P. Willett, "Combination of similarity rankings using data fusion," Journal of chemical information and modeling, vol. 53, pp. 1-10, 2013.

[30] G. Cano, J. García-Rodríguez, and H. Pérez-Sánchez, "Improvement of Virtual Screening Predictions using Computational Intelligence Methods," Letters in Drug Design & Discovery, vol. 11, pp. 33-39, 2014.