Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2015, 7(4):974-979

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

An algorithm for similarity-based virtual screening

Mubarak Himmat[1], Naomie Salim[1], Mohammed Mumtaz Al-Dabbagh[1] and Ali Ahmed[1,2]

[1]Faculty of computing, University of Technology Malaysia, Skudai, Johor, Malaysia
[2]Faculty of Engineering, Karary University, Khartoum, Sudan

---

ABSTRACT

*The virtual screening methods and techniques become one of the important and sophisticated ways of drug discovery, and molecules clustering, they are many methods proposed and applied in virtual screening, most of these screening methods used similarity coefficients to quantify the extent to which objects resemble one another. The result of using these similarity coefficients achieves a good result, and work is continuing to enhance and modify virtual screening methods or to present new methods, in this paper, we proposed a simple algorithm that uses simple equation that increased the effectiveness of virtual screening. The proposed method calculates the similarity and ranks the data to achieve better virtual screening results. We tested our proposed method with two benchmarks datasets Directory of Useful Decoys (DUD) data sets and Maximum Unbiased Validation (MUV) that already prepared and presented by 2D fingerprints ,the experiments have been conducted by selecting different 10 references from each activity class in each data sets , and we e evaluate the recall of active molecules at different at cut of 1% and 5% as usually done in virtual screening to evaluate the recall value, the overall results showed that the our proposed algorithm has good result in ligand-based virtual screening after comparing the result with Tanimoto coefficient which considered the standard similarity measure virtual screen*ing.

Keywords: Chemoinformatics; data fusion; similarity searching; Virtual screening; Drug Discovery**.**

INTRODUCTION

The process of discovering new drug is very difficult ,complex ,and costly ,computer technology and data mining tools have contributed in facilitating these process, nowadays ,virtual screening become one of the effective tools of drug discovery , there are a lot of ligand-based virtual screening and QSAR methods that have been proposed and applied in screening , there are some good reviews papers that discussed and covered these methods[1-4], ligand based virtual screening has there three different type of search ,structure searching, substructure search and similarity search ,in this article we will concern on similarity search where the search look to find the highly similar compounds in database that look like the references queries, many similarity measures have been applied and suggested in ligand based virtual screening [5-10]. Most of the similarity measure and techniques have been derived from the methods that already used in binary system and textual information retrieval or from documents retrieving methods ,the similarity searching here will look for all the structures in a database that are achieving the highly similar to a given structure. 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" is that the compounds with similar structures are likely to exhibit similar biological behaviours ,beside similarity coefficients , there are other methods that have been used in virtual screening like machine learning methods[11, 12]. Virtual screening as general has three different approaches, Ligand-based and structured-based Similarity searching ,the two approaches have been covered by researches ,but ligand-based VS on a chemical database has become widely used and many methods of information retrieval form different areas have been adapted an applied to ligand based virtual screening.in each process of similarity measure is core idea of the most methods is to find the best solution to measure and quantify the degree of similarity between the queries' reference structure' and each of the

molecules structures that stored in the database that screened. Virtual screening as general has three different approaches, Ligand-based and structured–based Similarity searching ,the two approaches have been covered by researches ,but ligand-based VS on a chemical database has become widely used and many methods of information retrieval form different areas have been adapted an applied to ligand based virtual screening.in each process of similarity measure is core idea of the most methods is to find the best solution of that measure and quantify the degree of similarity between the queries' reference structure' and each of the molecules structures that stored in the database that screened. The basic idea and method that applied for measuring these similarity is the similarity coefficients ,and there are many similarity coefficients that applied in virtual screening the work that done by[13, 14] discussed more than sixty similarity measures that used in Chemoinformatics and virtual screening ,and one of the standard similarity measure that gave good results and become standard similarity coefficient in virtual screening is Tanimoto coefficient ,but also there are other similarity measures that used by combination the ranking with different similarity measures concatenations and used together as fusion and resulted high performance, and there are many works that used fusion in virtual screening in early time of virtual screening research and recently[15-19] , till now the Tanimoto coefficient is considered the standard coefficients in molecules similarity searching .

EXPERIMENTAL SECTION

The proposed Method

Most of similarity measures are that are used in virtual screening are derived from Tanimoto coefficient, our proposed algorithm depends on a simple calculation method they can be used to calculate the similarity, and then we rank the molecules descending though that the highest molecules that similar to a query will be in the top, the proposed algorithms built into account only the mentioned following criteria:

1. The number of the only features that have and equal values.
2. The features that have value, but the values of the features do not equal each other.
3. The features that appear only in one molecule and none appear to others.
4. The total features of molecules.

The proposed algorithm calculates the similarity of the molecules by checking the molecules features by feature, and then calculating the total of each above mentioned criteria's, this proposed algorithm idea has been derived from the similarity measure that has been used in text and document retrieving area. The proposed algorithm depends on the presence and absence feature like other similarity coefficients but it relies especially on giving high consideration to the features that have non values (zero values), increases the percentage of the similarity between objects, and approved that ignoring (zero values) give good result of the recall. The new algorithm has been evaluated using benchmarked data which used in most of research on virtual screening, the conducted experiments show that the new proposed algorithm achieved good result comparing with Tanimoto that considered the standard virtual screening coefficient.

The proposed algorithm architecture

Part1 Calculating the similarity

BEGIN of the algorithm
For 1 to the total features number
Check feature by feature (until the end of the feature)
IF feature equal zeros
Count the count (count non presence value)
ELSE if
IF feature is equal and greater than zero Count
Count the absence features
IF feature of the query are not same, but greater than zero
Summation the fixed given value
Else
Count the absence features
END IF

Calculate the similarity using the proposed equation

Part 2 ranking the molecules descending

For 1: the number of molecules in datasets
Sort the similarity results in decreasing order
Calculate the 1% and 5% cut off
End

Each attribute of *m1* and *m*2 will be calculated, and the final equation will be as below
S_{m1m2} =absolute value (a +p)/ (n-c)

Each attribute of *m1 (query)* and *m*2 (reference) will be calculated, and the final equation will be as below
S_{m1m2}=abs ((a+p) /(n-c));

Where
a= the total number of features in both *m*1 and *m*2 that have and equal values.
p= fixed value (threshold =0. 05) for all the features m1 and *m2* that have value, but the values of the features do not equal to each other.
c= the number of features that have zero value in only *m*1 or *m*2.
N=the total number of features in each record of molecule (number of features).

METHOD

In this work we have proposed anew virtual screening ranking and screening algorithms ,the method is applied to For 2D ligand-based virtual screening, in our conducted experiments we have used two benchmarks ,the data sets have been chosen and used after developed and prepared by Scitegic's Pilot software [20], we applied in the methods to the both the MDDR (MDL Drug Data Report)[21], MUV dataset , the experiments are done by selection some references randomly for each process as for all activity classes . The results of the screening are compared to the Tanimoto considered as a reference standard in ligand-based VS of chemical datasets, and latest fusion methods that have been used proposed in virtual screening of ligand based virtual.

Dataset

In experiments we used the most popular and benchmarks datasets DUD which recently have been used in ligand based virtual screening and in duking methods[22, 23], the datasets firstly have been converted to Pipeline Pilot ECFC_4 (extended connectivity fingerprints and folded to size 1024 bits),the selected classes of DUD is shown on table 1, the all conducted experiments are done on the both datasets .The second dataset Is maximum unbiased validation MUV that considered one of the common datasets ,the data has seventeen different classes as shown on table 2, each class has 15030 molecules.

EXPERIMENTAL SECTION

In experiments, we conduct the screening process by applying the proposed algorithm investigate the effectiveness of using the algorithm , and the then compare the results with Tanimoto coefficient ,in the experiments we used ten references (queries)of active molecules for looking up to the similar molecules in the database, the measurement of the similarity between two molecules in VS is described by the degree of sharing features between these molecules. The work conducted on two datasets MUV and DUD that contain 2D chemical structure databases, The searches are carried out by selecting 10 references (active compounds) randomly to use as reference structures from each activity class and apply the algorithm to obtain the activity score for all of its compounds, to be fare we have used the ten 10 reference structures in all screening process for each activity classes. These processes have been applied to in same manner to the two datasets the results of screening have been evaluated by calculating the recall at 1% and 5%. Here 1% and 5% represent the percentage numbers of the databases molecules belonging to the same activity classes as the reference structure that is retrieved in the top 1% and 5% of a ranking of the databases. For the data set MUV ,there are seventeen activity classes are tested, the screening have been conducted by applying and 10 active compounds (references)randomly to use as reference structures, and the then measure obtains the activity score for all activity class compounds.

RESULTS AND DISCUSSION

The results of all conducted experiments that use DUD and MUV datasets that were obtained a good and high resulted in Table 3 and Table 4 .The screening results demonstrated that our new algorithm achieved better results when they were compared by Tanimoto, and these results achieved showed the efficiency of the new algorithms, the both tables contains the activity classes and recall values and average of the recall values,. The first column represents the activity class of the dataset, the next column represents the average recall obtained from the top 1% and the top 5% ranking for each of the activity classes by using Tanimoto coefficient and two last two columns showed the proposed algorithm results , the last raw of the tables showed the overall rankings results 1% and 5% average results, the mean recalls, we found that our method achieved a clear and good result in 5% of recall and a little bit results in 1% recalls, the result and also the results that obtained by screening process of MUV dataset achieved significant results for it gave results that increased more than two point and this could be consider high

results one compared with Tanimoto the standard coefficient in virtual screening ,And actually the experiments showed and confirmed that the , the new proposed method is presented enhancement in virtual screening.

Table 1.DUD Selected 11 activity classes

Activity class	Active molecules
FGFR1T	120
FXA	146
GART	40
GBP	52
GR	78
HIVPR	62
HIVRT	43
HMGA	35
HSP90	37
MR	15
NA	49
PR	27

Table 2 .MUV activity classes

Activity index	Activity class
466	S1P1 rec. (agonists)
548	PKA (inhibitors)
600	SF1 (inhibitors)
644	Rho-Kinase2 (inhibitors)
652	HIV RT-RNase (inhibitors)
689	Eph rec. A4 (inhibitors)
692	SF1 (agonists)
712	HSP 90 (inhibitors) 30
713	ER-a-Coact. Bind. (inhibitors)
733	ER-b-Coact. Bind. (inhibitors)
737	ER-a-Coact. Bind. (potentiators)
810	FAK (inhibitors
832	Cathepsin G (inhibitors)
846	FXIa (inhibitors)
852	FXIIa (inhibitors)
858	D1 rec. (allosteric modulators)
859	M1 rec. (allosteric inhibitors)

Table 3. The recall is calculated using the top 1% and top 5% of the DUD 12 selected activity classes

Selected activity	Tanimoto		Proposed Method	
Classes	**1%**	**5%**	**1%**	**5%**
Fgfr1t	2.97	7.25	**3.167**	**7.41**
Fxa	3.01	8.7	**2.74**	**9.58**
Gart	5.4	21.25	**7**	**23.25**
Gbp	15.16	27.12	**16.154**	**28.07**
Gr	2.45	6.79	**2.82**	**8.205**
Hivpr	5.22	12.74	**4.35**	**13.871**
Hivrt	2.31	4.42	**1.86**	**6.511**
Hmga	5.45	10.02	**5.714**	**8.571**
Hsp90	3.58	9.46	**3.24**	**10.54**
Mr	3.87	8.67	**4.667**	**10**
Na	3.04	6.53	**3.878**	**6.11**
Pr	1.53	5.93	**2.96**	**7%**
Avg	**4.5**	**10.74**	**4.88**	**11.015**

Table 4. The recall is calculated using the top 1% and top 5% of the MUV activity classes

Activity index	Tanimoto		Proposed Method	
	1%	5%	**1%**	**5%**
466	3.1	5.86	**3**	**7.33**
548	8.62	22.76	**12.67**	**24**
600	3.79	11.38	**4**	**10.33**
644	7.59	17.59	**8**	**17.67**
652	2.76	7.93	**2.67**	**8.33**
689	3.79	9.66	**3.33**	**7.33**
692	0.69	4.83	**2**	**8**
712	4.14	10.34	**2.67**	**7.33**
713	3.1	7.24	**2.67**	**5.67**
733	3.45	8.97	**2.67**	**6.33**
737	2.41	8.28	**2**	**9.43**
810	2.07	6.9	**1.67**	**8.67**
832	6.55	13.1	**8.67**	**17.67**
846	9.66	28.62	**9.67**	**25.67**
852	12.41	21.38	**11.33**	**22**
858	1.72	5.86	**2.33**	**5**
859	1.38	8.97	**1.67**	**7.67**
Average	4.54	11.70	**4.765882**	**11.67**

CONCLUSION

This study proposed a new method of virtual screening, the proposed method are tested using two benchmarks datasets, the proposed algorithm concepts are adapted by give high consideration to features that have non values (zero values) and this increases the percentage of the similarity between molecules, the results that obtained by the conducted experiments on DUD and MUV achieved good enhancement in ligand based virtual screening, the results are compared with Tanimoto results. By the obtained results we recommended using of our proposed similarity algorithm in virtual screening of 2D fingerprint chemical database, the overall a obtained results of the proposed method showed that clearly the screening search using our algorithm outweighed the Tanimoto results, The significance test conducted on that VS results of all datasets showed that good enhancement has happen achieved.

Acknowledgments
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] W.P. Walters, M.T. Stahl, M.A. Murcko, *Drug Discovery Today*, 3 (**1998**) 160-178.
[2] P. Willett, *Annual review of information science and technology*, 43 (**2009**) 1-117.
[3] H. Geppert, M. Vogt, J.r. Bajorath, *Journal of chemical information and modeling*, 50 (**2010**) 205-216.
[4] M.A. Johnson, G.M. Maggiora, Concepts and applications of molecular similarity, Wiley, **1990**.
[5] G.M. Downs, P. Willett, W. Fisanick, *Journal of Chemical Information and Computer Sciences*, 34 (**1994**) 1094-1102.
[6] G.M. Downs, P. Willett, *Reviews in computational chemistry*, 7 (**1996**) 1-66.
[7] P. Willett, *Biochemical Society Transactions*, 31 (**2003**) 603-606.
[8] P. Willett, Drug discovery today, 11 (**2006**) 1046-1053.
[9] T. Girschick, L. Puchbauer, S. Kramer, *Journal of cheminformatics*, 5 (**2013**).
[10] N. Salim, J. Holliday, P. Willett, *Journal of chemical information and computer sciences*, 43 (**2003**) 435-442.
[11] J. Chen, J. Holliday, J. Bradshaw, *Journal of chemical information and modeling*, 49 (**2009**) 185-194.
[12] J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, *Journal of chemical information and modeling*, 46 (**2006**) 462-470.
[13] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, P. Willett, *Journal of chemical information and modeling*, 52 (**2012**) 2884-2901.
[14] V. Consonni, R. Todeschini, *Match-Communications in Mathematical and Computer Chemistry*, 68 (**2012**) 581.
[15] P. Willett, *Journal of chemical information and modeling*, 53 (**2013**) 1-10.
[16] P. Willett, *Computational and Structural Biotechnology Journal*, 5 (**2013**).
[17] F. Svensson, A. Karlén, C. Sköld, *Journal of chemical information and modeling*, 52 (**2011**) 225-232.
[18] B. Chen, C. Mueller, P. Willett, *Molecular Informatics*, 29 (**2010**) 533-541.
[19] P. Willett, *QSAR & Combinatorial Science*, 25 (**2006**) 1143-1152.

[20] S.P. Pilot, Inc.: San Diego, CA, (**2008**).[21]
[22] A. Ahmed, F. Saeed, N. Salim, A. Abdo, *molecules*, 1 (**2014**) 2.
[23] N. Huang, B.K. Shoichet, J.J. Irwin, *Journal of medicinal chemistry*, 49 (**2006**) 6789-6801.