Websites:
http://www.sciencepub.net
http://www.sciencepub.net/researcher

Emails:
editor@sciencepub.net
marslandresearcher@gmail.com

Researcher

MARSLAND PRESS
Multidisciplinary Academic Journal Publisher

Fragment Selection and Reweighting Factor Combination Method for Ligand-based Virtual Screening

Ali Ahmed A.Abdelrahim[1,2], Naomie Salim[1] and Ammar Abdo[1]

[1]Soft Computing Research Group, Faculty of Computing, Universiti Tecknologi Malaysia, 81310 Skudai, Malaysia
[2]Faculty of Engineering, Karary University, Khartoum 12304, Sudan
alikarary@gmail.com

Abstract: Many methods have been developed to capture biological similarities between two compounds to aid in the discovery of new pharmaceuticals. Of the variety of similarity metrics that have been introduced, the Tanimoto coefficient and Bayesian networks are the most prominent. Recently, the use of the Bayesian network, as an alternative to existing tools for similarity-based virtual screening, has received noticeable attention from researchers in the chemoinformatics field. In our previous works, the retrieval performance of the Bayesian network was observed to improve significantly when multiple reference structures or relevance feedback information were used. In this article, the authors enhance the Bayesian inference network (BIN) using combination similarity method. In this approach, the important fragments were filtered from the molecular fingerprint fragments, and then a fragment reweighting process was used to reformulate the weights of the selected fragments. At the end, our simulated virtual screening experiments with MDL Drug Data Report data sets showed that this approach significantly improved the retrieval effectiveness of ligand-based virtual screening, especially when the active molecules being sought have a high degree of structural heterogeneity.

[A.Abdelrahim A Ali, Salim N, Abdo A. **Taxonomic Fragment Selection and Reweighting Factor Combination Method for ligand-based Virtual Screening**. *Researcher* 2021;13(8):37-47]. ISSN 1553-9865 (print); ISSN 2163-8950 (online). http://www.sciencepub.net/researcher.7. doi:10.7537/marsrsj130821.07.

Keywords: Chemoinformatics; virtual screening; fragment selection; fragment reweighting

1. Introduction

Many virtual screening (VS) approaches have been implemented for searching chemical databases, such as substructure search, similarity, docking and QSAR. Of these, similarity searching is the simplest, and one of the most widely-used techniques, for ligand-based virtual screening (LBVS) [1].

Virtual screening refers to the use of a computer-based method to process compounds from a library or database of compounds in order to identify and select ones that are likely to possess a desired biological activity, such as the ability to inhibit the action of a particular therapeutic target. The selection of molecules with a virtual screening algorithm should yield a higher proportion of active compounds, as assessed by experiment, relative to a random selection of the same number of molecules [2].

There are many studies in the literature associated with the measurement of molecular similarity [1, 3-6]. The most common approach, which we used in this study, characterises molecules using 2D fingerprints that encode the presence of 2D fragment substructures in a molecule. The 2D fingerprints involve the specification of the entire structure of a molecule. This specification is generated for both the ligand molecule and each molecule in the database. The similarity between the ligand molecule and the molecule in the database is then computed using the number of substructural fragments they have in common and an association coefficient such as the Tanimoto coefficient [1, 7].

Recent works have suggested that significant improvements in retrieval effectiveness can be achieved by combining results from multiple similarity coefficients, multiple reference structures and multiple molecular descriptors [8, 9]. Recently, the Bayesian inference network model has been introduced for performing molecular similarity searching [10, 11]. One of the most important characteristics of the inference network model is that it enables the combination of more than one enhancement technique. In our previous works, the retrieval performance of Bayesian inference network was observed to improve significantly when relevance feedback and turbo search screening were used [10].

Features Selection (FS) is a process of selecting a subset of features available from the data for the application of a learning algorithm. The best feature subset is one that contains the least number of

features that most contribute to accuracy and efficiency. This is an important stage of preprocessing and is one of the two ways of avoiding high dimensional space of features; the other is feature extraction. The elimination of the unimportant and obsolete features results in enhances and improves the recall and classification rate [12, 13]. The current molecule's fingerprint consists of many fragments or features, not all of which have the same importance, and removing the unimportant fragments can enhance the recall of similarity measures [14]. Recent studies, based on the same data sets of this study, used feature selection to enhance the molecular similarity were found in [15, 16].

Fragment reweighting is the process of assigning higher weights to the fragments that occur more frequently in the set of active reference structures, while others are penalized. Fragment or feature reweighting is one of the most useful query modification techniques in IR systems [17-20]. In our previous works, the retrieval performance of the Bayesian inference network was observed to improve significantly when ligand expansion was used [21].

In this study, we proposed the combination approach of reweighted selected fragments to enhance the screening electiveness of Bayesian inference network (BIN). In this approach, the supervised statistical feature selection algorithm was applied first to determine the important fragments which showed a strong correlation with the class identifier. The output of this stage is used as the input for the reweighting process, in which the reweighting factors were calculated and used to reformulate the weights of the selected fragments, resulting in reweighted selected fragments.

2. Material and Methods

This study has compared the retrieval results obtained using three different similarity-based screening models. The first screening system was based on the Tanimoto (TAN) coefficient, which has been used in ligand-based virtual screening for many years and is now considered a reference standard. The second model was based on a basic BIN [11] using the Okapi (OKA) weight, which was found to perform the best in the experiments and which we shall refer to as the conventional BIN model [22]. The third model, which is our proposed model, is a BIN based on the combination approach of Reweighted Selected Fragments, which we shall refer to as the BINRSF model. In the following paragraphs, we give a brief description of each of these three models.

2.1 Tanimoto-based Similarity Model

This model used the continuous form of the Tanimoto coefficient, which is applicable to the non-binary data of the fingerprint. $S_{K,L}$ is the similarity between objects or molecules K and L, which, using Tanimoto, is given by Eq. 1:

$$S_{KL} = \frac{\sum_{j=1}^{M}(w_{jk}w_{jl})}{\sum_{j=1}^{M}(w_{jk})^2 + \sum_{j=1}^{M}(w_{jl})^2 - \sum_{j=1}^{M}(w_{jk}w_{jl})} \quad (1)$$

For molecules described by continuous variables, the molecular space is defined by an M × N matrix, where entry w_{ji} is the value of the jth fragments ($1 \le j \le M$) in the ith molecule ($1 \le i \le N$). The origins of this coefficient can be found in a review paper by Ellis et al. [23].

2.2 Conventional BIN Model

The conventional BIN model, as shown in Fig. 1, is used in molecular similarity searching. It consists of three types of nodes: compound nodes as roots, fragment nodes and a reference structure node as leaf. The roots of the network are the nodes without parent nodes and the leaves are the nodes without child nodes. Each compound node represents an actual compound in the collection and has one or more fragment nodes as children. Each fragment node has one or more compound nodes as parents and one reference structure node as a child (or more where multiple references are used). Each network node is a binary value, taking one of the two values from the set {true, false}. The probability that the reference structure is satisfied given a particular compound is obtained by computing the probabilities associated with each fragment node connected to the reference structure node. This process is repeated for all the compounds in the database. The resulting probability scores are used to rank the database in response to a bioactive reference structure in the order of decreasing probability of similar bioactivity to the reference structure.

To estimate the probability associating each compound to the reference structure, the probability for the fragment and reference nodes must be computed. One particular belief function, called OKA, has been found to have the most effective recall [22]. This function is used to compute the probabilities for the fragment nodes and is given by Eq. 2:

$$bel_{OKA1}(f_i) = \alpha + (1-\alpha) \times \frac{ff_{ij}}{ff_{ij} + 0.5 + 1.5 \times \frac{|C_j|}{|C_{avg}|}} \times \frac{\log\left|\frac{m + 0.5}{cf_j}\right|}{\log(m+1)} \times \frac{\min(ff_{ij}, ff_{ir})}{\max(ff_{ij}, ff_{ir})} \quad (2)$$

where: α is a constant, and experiments using the Bayesian network show that the best value is 0.4 [11, 24], ff_{ij} and ff_{ir} are the frequency of the ith fragment within jth compound and r reference structure respectively; cf_i is the number of compounds containing ith fragment; |cj| is the size (in terms of the number of fragments) of the jth compound; $|C_{avg}|$ is the average size of all the compounds in the database; and m is the total number of compounds.

To produce a ranking of the compounds in the collection with respect to a given reference structure, a belief function from In Query, the SUM operator, was used. If p1, p2,..., pn represents the belief in the fragment nodes (parent nodes of r), then the belief at r is given by Eq. 3:

$$bel_{sum}(r) = \frac{\sum_{i=1}^{n} p_i}{n} \quad (3)$$

where n is the number of the unique fragments assigned to r reference structure, and pi is the value of the belief function bel(fi) in ith fragment node.

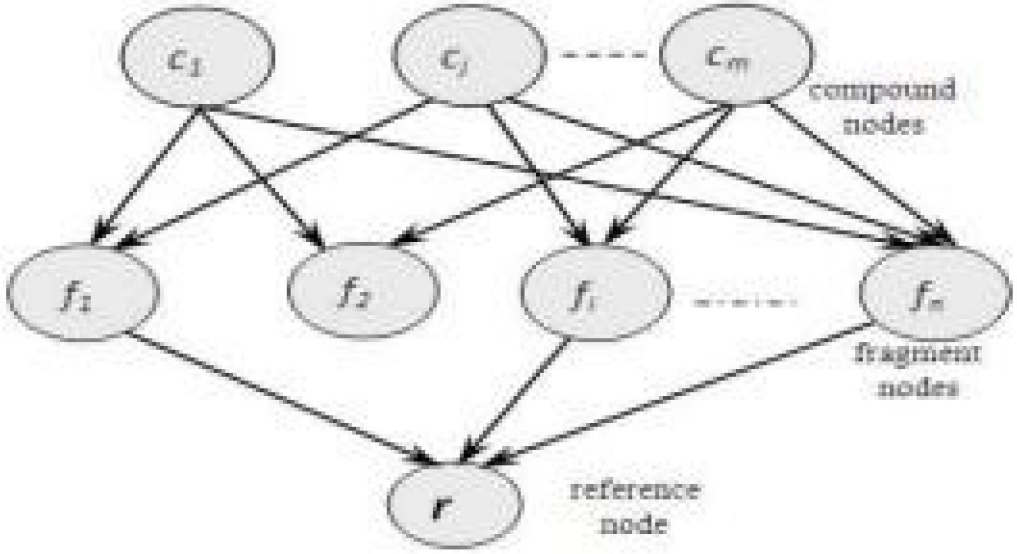


Figure 1: Bayesian inference network model.

2.3 Combination Similarity Model

2.3.1 Fragment selection stage

In this stage, the supervised statistical feature selection algorithm was applied to determine the important fragments which showed a strong correlation with the class identifier. This algorithm considered one attribute at a time to see how well each predictor alone (fragment) predicted the target variable (output). The importance value of each variable was then calculated as (1- p) where p is the association strength between the candidate predictor (fragment) and the target variable. Since the target values were continuous, p values was based on the F statistic.

Part of the output of the feature selection algorithm for the first class of DS1 data set (details of these data sets are shown in the experimental design section) is illustrated in Table 1. In this table, the fragments were ranked according to their importance. The predictors were then labelled as 'important', 'marginal' and 'unimportant' for values above 0.95, between 0.95 and 0.90, and below 0.90 respectively. Only the 'important' fragments, from the above table, were used and the 'marginal' and 'unimportant' fragments were ignored before executing the search process, similar study based on this algorithm found in [25].

2.3.2 Fragment reweighting stage

In the reweighting process stage, the reweighting factor rwfi was calculated for each reference i of the input references using the following equation:

$$rwf_i = \frac{SF_{fi}}{\max SF} \quad (4)$$

where S_{Ffi} is the frequency of selected ith fragment in the set of references' input and maxSF is the maximum frequency of selected fragment in the set of references inputs.

Table 1: Part of the output of fragment selection

Rank	Fragment	Type	Importance	Value
1	F931	Range	Important	1.0
2	F586	Range	Important	1.0
3	F546	Range	Important	1.0
.	.	.	.	.
130	F484	Range	Important	0.96
131	F609	Range	Marginal	0.94
132	F485	Range	Marginal	0.93
133	F522	Range	Marginal	0.92
.	.	.	.	.
248	F52	Range	Unimportant	0.0
249	F742	Range	Unimportant	0.0
250	F775	Range	Unimportant	0.0

2.3.3 Combination stage

A new reweighted ligand RLi was formed by adding the original weight of each selected fragment to the reweighting factor of each reference based on the following equation:

$$nsw_i = sw_i + rwf_i \quad (5)$$

where sw_i is the original frequency of the selected ith fragment in the reference input.

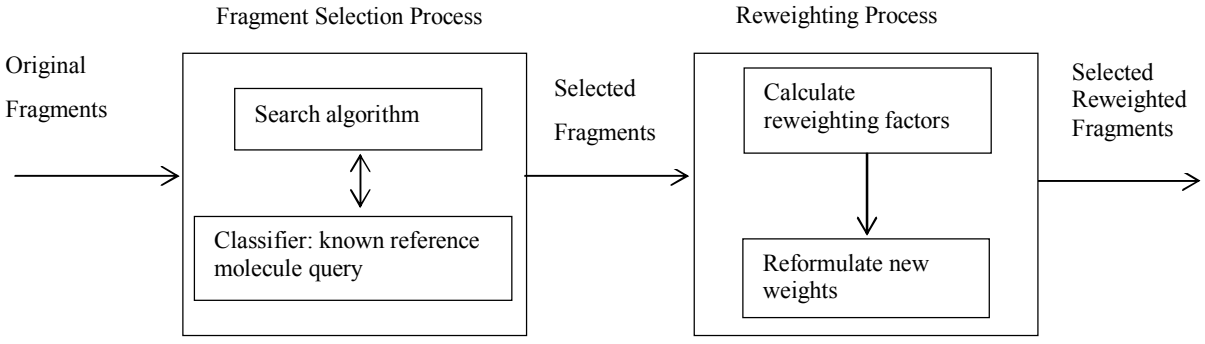


Figure 2: Combination of fragment selection and reweighting process

Consequently, the use of (4) and (5) to assign the new weights shows that higher weights will be assigned to those that occur more frequently in the set of references' input structures. The simplified diagram of this approach is illustrated in Fig. 2.

3. Experimental design

The searches were carried out using the most popular chemoinformatics database: the MDL Drug Data Report (MDDR) [26], with 102,516 molecules. All molecules in the MDDR database were converted to Pipeline Pilot ECFC4 (extended connectivity fingerprints and folded to size 1024 bits) [27]; this data set has been used recently by our research group in our previous study[28].

For screening experiments, three data sets (DS1, DS2 and DS3) were chosen from the MDDR database. Data set DS1 contains 11 MDDR activity classes, with some of the classes involving actives that are structurally homogeneous, and others involving actives that are structurally heterogeneous (structurally diverse). The DS2 data set contains 10 homogeneous MDDR activity classes and the DS3 data set contains 10 heterogeneous MDDR activity classes. Full details of DS1 data sets are given in Tables 2. Each row in the tables contains an activity class, the number of molecules belonging to the class and the class' diversity, which was computed as the mean pair-wise similarity calculated across all pairs of molecules in each class. These pair-wise similarity calculations for all data sets were performed using Pipeline Pilot software [27]. For each data set (DS1-DS3), the screening experiments were conducted with 10 reference structures selected randomly from each activity class and the similarity measure used to obtain an activity score for all of its compounds. These activity scores were then sorted in descending order with the recall of the active compounds, meaning the percentage of the desired activity class compounds that are retrieved in the top 1% and 5% of the resultant sorted activity scores, provide a measure of the performance of our similarity method.

4. Results and Discussion

Our goal is to identify the different retrieval effectiveness of using different search approaches. In this study, we tested the TAN, BIN, BINRSF models against the MDDR database using three different data sets (DS1-DS3).

Table 2: MDDR structure activity classes for DS1 data set

Activity index	Activity Class	Active molecules	Pair-wise similarity (mean)
31420	Renin inhibitors	1130	0.290
71523	HIV protease inhibitors	750	0.198
37110	Thrombin inhibitors	803	0.180
31432	Angiotensin II AT1 antagonists	943	0.229
42731	Substance P antagonists	1246	0.149
06233	Substance P antagonists	752	0.140
06245	5HT reuptake inhibitors	359	0.122
07701	D2 antagonists	395	0.138
06235	5HT1A agonists	827	0.133
78374	Protein kinase C inhibitors	453	0.120
78331	Cyclooxygenase inhibitors	636	0.108

The results of the searches of DS1-DS3 are presented in Tables 3-5 respectively, using cut offs at both 1% and 5%. In these tables, the first column from the left contains the results for the TAN, the second column contains the corresponding results when BIN is used and the last column of each table contains the corresponding results when BINRSF is used. Each row in the tables lists the recall for the top 1% and 5% of a sorted ranking when averaged over the ten searches for each activity class. The mean rows in the tables correspond to the mean when averaged over all activity classes, and the CI rows represent the 95% confidence interval. The similarity method with the best recall rate in each row is strongly shaded, and the best mean recall value is boldfaced. The bottom row in a table corresponds to the total number of shaded cells for each similarity method across the full set of activity classes.

Table 3: Retrieval results calculated using top 1% and 5% for DS1 data set using TAN, BIN, and BINRSF

Activity Index		1%			5%		
		TAN	BIN	BINRSF	TAN	BIN	BINRSF
31420		55.84	74.08	85.02	85.49	87.61	98.15
71523		22.26	28.26	47.97	42.7	52.72	69.47
37110		12.54	26.05	46.9	24.11	48.2	75.29
31432		33.36	39.23	45.9	68.2	77.57	91.04
42731		16.24	21.68	33.67	32.81	26.63	49.73
06233		14.23	14.06	21.45	27.01	23.49	36.3
06245		10.06	6.31	3.32	22.9	14.86	12.57
07701		8.91	11.45	22.01	23.1	27.79	46.17
06235		11.87	10.84	18.2	24.54	23.78	43.84
78374		16.75	14.25	24.71	24.26	20.2	41.35
78331		8.05	6.03	9.26	16.83	11.8	17.73
Mean		19.10	22.93	32.58	35.63	37.69	52.88
CI	Lower	9.59	9.64	17.15	21.01	20.52	34.25
	Upper	28.61	36.22	48.01	50.25	54.86	71.50
Shaded cells		1	0	10	1	0	10

Table 4: Retrieval results calculated using top 1% and 5% for DS2 data set using TAN, BIN, and BINRSF

Activity Index		1%			5%		
		TAN	BIN	BINRSF	TAN	BIN	BINRSF
07707		78.3	72.18	72.67	91.08	74.81	77.57
07708		74.01	96	98.12	88.52	99.61	100
31420		46.44	79.82	97.21	77.6	95.46	97.66
42710		57.22	76.27	98.64	67.59	92.55	99.45
64100		93.22	88.43	90.35	97.89	99.22	99.64
64200		63.39	70.18	79.77	89.82	99.2	99.65
64220		73.56	68.32	84.33	92.05	91.32	99.32
64500		60.75	81.2	93.64	74.98	94.96	99.81
64350		76.69	81.89	94.92	90.34	91.47	99.62
75755		95.99	98.06	98.91	98.78	98.33	98.54
Mean		71.95	81.235	90.86	86.86	93.69	97.14
CI	Lower	60.86	73.89	84.37	79.59	88.41	92.18
	Upper	83.05	88.59	97.34	94.14	98.97	99.50
Shaded cells		2	0	8	2	0	8

Table 5: Retrieval results calculated using top 1% and 5% for DS3 data set using TAN, BIN, and BINRSF

Activity Index		1%			5%		
		TAN	BIN	BINRSF	TAN	BIN	BINRSF
09249		25.09	15.33	25.12	40.21	25.72	39.56
12455		7.7	9.37	9.11	19.08	14.65	12.29
12464		9.02	8.45	16.11	14.56	16.55	36.71
31281		27.53	18.29	49.33	44	28.29	62.86
43210		11.1	7.34	10.05	26.37	14.41	19.61
71522		2.35	4.08	8.87	6.28	8.44	20.67
75721		24.02	20.41	34.9	28.97	30.02	61.32
78331		6.27	7.51	11.07	15.79	12.03	19.51
78348		4.69	9.79	10.18	13.16	20.76	27.54
78351		4.31	13.68	16.74	10.55	12.94	21.21
Mean		12.21	11.42	19.15	21.90	18.38	32.13
CI	Lower	5.37	7.65	9.48	12.84	13.07	19.39
	Upper	19.05	15.19	28.8	30.95	23.69	44.87
Shaded cells		1	1	8	3	0	7

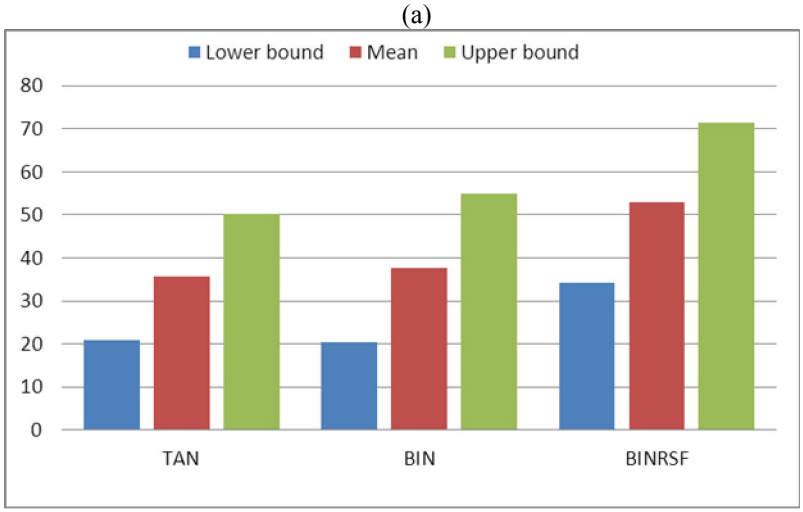
Table 6: Rankings of TAN, BIN, BINRSF approaches Based on Kendall W Test Results: DS1-DS3 at top 1% and top 5%

Data set	Recall type	W	P	Ranking
DS1	1%	0. 504	0.004	BINRSF>BIN>TAN
	5%	0.502	0.003	BINRSF>TAN>BIN
DS2	1%	0.490	0.007	BINRSF>BIN>TAN
	5%	0.500	0.006	BINRSF>BIN>TAN
DS3	1%	0.480	0.008	BINRSF>TAN>BIN
	5%	0.495	0.009	BINRSF>BIN>TAN

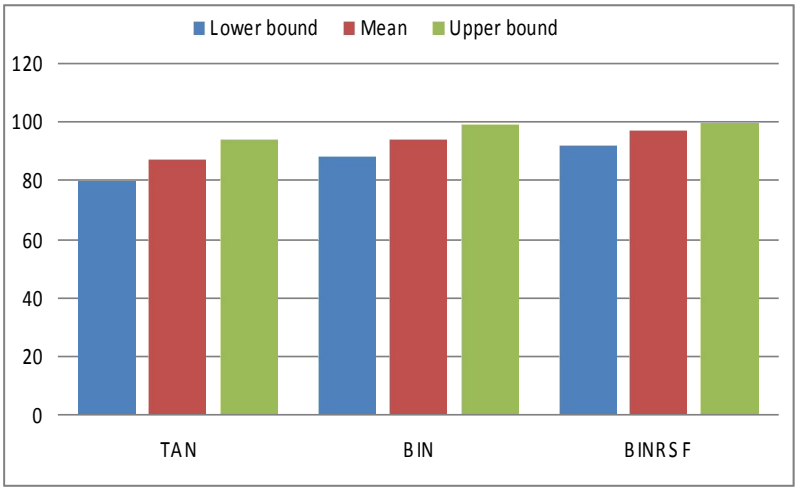
A look at the recall values in Tables 5 to 7 enables comparisons to be made between the effectiveness of the various search models. However, a more quantitative approach is possible using the Kendall W test of concordance [29]. This test shows whether a set of judges make comparable judgments about the ranking of a set of objects. Here, the activity classes were considered the judges and the recall rates of the various search models, the objects. The outputs of this test are the value of the Kendall coefficient and the associated significance level, which indicates whether the value of the coefficient could have occurred by chance. If the value is significant (for which we used cutoff values of both 0.01 and 0.05), then it is possible to give an overall ranking of the objects that have been ranked.

The results of the Kendall analyses for DS1, DS2 and DS3 are reported in Table 6 which describes the top 1% and top 5% rankings for the various searching approaches. In this Table, the columns show the data set type, the recall percentage, the value of the Kendall's coefficient of Concordance (W), the associated probability (p) and the ranks of each of the different searching methods. Table 6 shows that the values of Kendall coefficients vary from 0.504 (agreement is 50.4%) for DS1 (top 5%) to 0.48 (agreement is 48%) for DS3 (top 1%) while the values of associated probability, (p), is (<0.01) for all recall percentages of the three data sets. This indicates that these values are significant and it became possible to give an overall ranking to the objects (searching approaches).

(a)



(b)



(c)

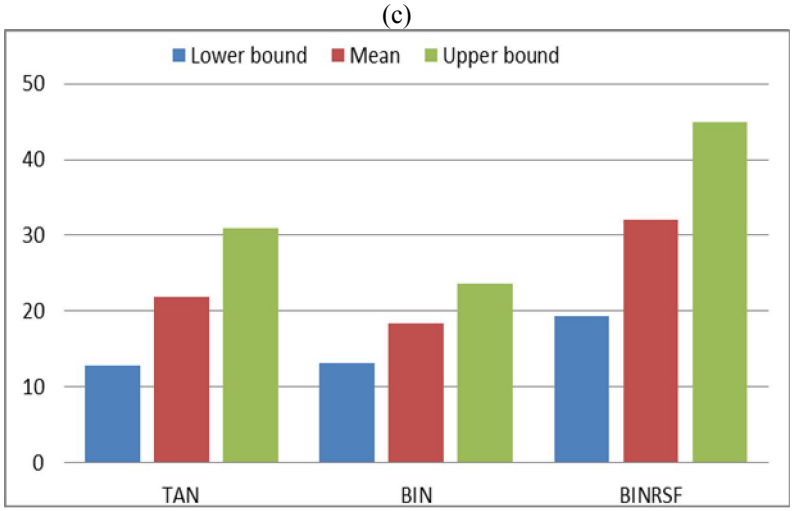


Figure 3: Performance with 95% confidence bound for the three screening methods with a) DS1, b) DS2 and c) DS3 data sets at top 5%

Therefore, the ranking of the search methods for all three cases is significant and has not occurred by chance.

Some of the activity classes, such as low-diversity activity classes, may contribute disproportionately to the overall value of mean recall. Therefore, using the mean recall value as the evaluation criterion could be impartial in some methods, but not in others. To avoid this bias the effective performances of the different methods have been further investigated based on the total number of shaded cells for each method across the full set of activity classes. This is shown in the bottom rows of Tables 5 to 7. These shaded cell results are also listed in Table 7. According to the total number of shaded cells in this table, BINRSF was the best performing search across the three data sets DS1-DS3.

The results of the DS1 search shown in Table 3 show that BINRSF produced the highest mean value compared with other measures. The values of the Kendall coefficient for DS1 (top1% and 5%) are 0.504 and 0.502 respectively. Given that the results are significant, since associated probability is <0.01, the overall ranking of the different approaches is BINRSF>BIN>TAN and for the top 1%, which shows that the combination method has a high rank value.

Table 7: Number of shaded cells for mean recall of actives using different search methods for DS1-DS3 Top 1% and 5%

Data Set	TAN	BIN	BINRSF
Top 1%			
DS1	1	0	10
DS2	2	0	8
DS3	1	1	8
Top 5%			
DS1	1	0	10
DS2	2	0	8
DS3	3	0	7

Similarly, the overall ranking of the different searching approaches for this data set at the top 5% is BINRSF>TAN>BIN, which again shows the best effectiveness of the combination method. For DS2 data set, the combination method (BINRSF) has the highest rank for both top 1% and top 5%. The DS3 searches are of particular interest, since they involve the most heterogeneous activity classes in the three data sets used, and thus provide a complete test of the effectiveness of a screening method. Tables 5-9 show that BINRSF gives the best performance out of all the methods for this data set at both cutoffs.

Fig. 3 showing the mean, lower and upper bounds of the confidence intervals of different methods, reveals that we can be 95% confident that the combination method (BINRSF) performs best for the DS1, DS2 and DS3 data sets. Therefore, on the basis of these results, we can say with 95% statistical certainty that the combination method search will do better than conventional similarity systems.

5. Conclusion

In this study, we have developed a combination method to enhance the effectiveness of Bayesian inference networks. This method is based on the combination of the two methods; the first method was the fragment selection method, in which important fragments were filtered from the molecular fingerprint fragment based on the supervised features selection approach. Secondly, the fragment reweighting method, which was based on the reweighting factor, was used as a second method to reformulate the weights of the selected fragments produced by the previous stage.

The overall results of this combination show that the screening similarity search of this method significantly outperformed the Tanimoto and conventional Bayesian inference networks similarity methods. In addition, there was evidence to suggest that the reweighting combination similarity method was more effective for high diversity data sets.

Acknowledgment

This study is supported by Research Management Centre (RMC) at Universiti Teknologi Malaysia under Post-Doctoral Fellowship Scheme for the Project: “Enhancing Molecular Similarity Search”.

Conflict of interest

The authors have no conflict of interest to declare.

References

[1]. Willett, P., J.M. Barnard, and G.M. Downs, Chemical similarity searching. Journal of Chemical Information and Computer Sciences, 1998. 38(6): p. 983-996.

[2]. Johnson, M.A. and G.M. Maggiora, Concepts and applications of molecular similarity. 1990: Wiley New York.

[3]. Sheridan, R.P. and S.K. Kearsley, Why do we need so many chemical similarity search methods? Drug discovery today, 2002. 7(17): p. 903-911.

[4]. Nikolova, N. and J. Jaworska, Approaches to measure chemical similarity–a review. QSAR & Combinatorial Science, 2003. 22(9‐10): p. 1006-1026.

[5]. Bender, A. and R.C. Glen, Molecular similarity: a key technique in molecular informatics. Organic & biomolecular chemistry, 2004. 2(22): p. 3204-3218.

[6]. Maldonado, A.G., et al., Molecular similarity and diversity in chemoinformatics: from theory to applications. Molecular diversity, 2006. 10(1): p. 39-79.

[7]. Gasteiger, J., Handbook of chemoinformatics, 2003.

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

[9]. Hert, J., et al., Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. Journal of chemical information and computer sciences, 2004. 44(3): p. 1177-1185.

[10]. Abdo, A., N. Salim, and A. Ahmed, Implementing relevance feedback in ligand-based virtual screening using Bayesian inference network. Journal of biomolecular screening, 2011. 16(9): p. 1081-1088.

[11]. [Abdo, A. and N. Salim, Similarity‐Based Virtual Screening with a Bayesian Inference Network. ChemMedChem, 2009. 4(2): p. 210-218.

[12]. Uzer, M.S., N. Yilmaz, and O. Inan, Feature Selection Method Based on Artificial Bee Colony Algorithm and Support Vector Machines for Medical Datasets Classification. The Scientific World Journal, 2013.

[13]. Zheng, Z., R. Srihari, and S. Srihari. A feature selection framework for text filtering. in Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. 2003.

[14]. Vogt, M., A.M. Wassermann, and J. Bajorath, Application of Information—Theoretic Concepts in Chemoinformatics. Information, 2010. 1(2): p. 60-73.

[15]. Bender, A., et al., Molecular similarity searching using atom environments, information-based feature selection, and a naive Bayesian classifier. Journal of Chemical Information and Computer Sciences, 2004. 44(1): p. 170-178.

[16]. [Bender, A., et al., Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. Journal of Chemical Information and Computer Sciences, 2004. 44(5): p. 1708-1718.

[17]. Haines, D. and W.B. Croft. Relevance feedback and inference networks. in Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. 1993. ACM.

[18]. De Campos, L.M., J.M. Fernández‐Luna, and J.F. Huete, Implementing relevance feedback in the Bayesian network retrieval model. Journal of the American Society for Information Science and Technology, 2003. 54(4): p. 302-313.

[19]. Xin, J. and J.S. Jin. Relevance feedback for content-based image retrieval using Bayesian network. in ACM International Conference Proceeding Series. 2004.

[20]. Kołcz, A. and C. Teo. Feature weighting for improved classifier robustness in CEAS'09: sixth conference on email and anti-spam. 2009.

[21]. Abdo, A., F. Saeed, H. Hamza, A. Ahmed and N. Salim, Ligand expansion in ligand-based virtual screening using relevance feedback. Journal of computer-aided molecular design, 2012. 26(3): p. 279-287..

[22]. Abdo, A. and N. Salim, New fragment weighting scheme for the Bayesian inference network in ligand-based virtual screening. Journal of chemical information and modeling, 2010. 51(1): p. 25-32.

[23]. Ellis, D., J. Furner-Hines, and P. Willett, Measuring the degree of similarity between

objects in text retrieval systems. Perspectives in Information Management, 1993. 3(2): p. 128-149.
[24]. Chen, B., C. Mueller, and P. Willett, Evaluation of a Bayesian inference network for ligand-based virtual screening. Journal of cheminformatics, 2009. 1(1): p. 1-10.
[25]. Bijanzadeh, E., Y. Emam, and E. Ebrahimie, Determining the most important features contributing to wheat grain yield using supervised feature selection model. Australian Journal of crop science, 2010. 4: p. 402-407.
[26]. Symyx Technologies. MDL drug datareport.http://www.symyx.com/products/ databases/bioactivity/mddr/index.jsp. Accessed October 20, 2011
[27]. Pipeline Pilot, Accelrys Software Inc.: San Diego, CA, (2008).
[28]. Ahmed, A., A. Abdo, and N. Salim, Ligand-based virtual screening using Bayesian inference network and reweighted fragments. The Scientific World Journal, 2012.
[29]. Siegel, S., Nonparametric statistics for the behavioral sciences. 1956.

2/5/2021