J Comput Aided Mol Des (2012) 26:279–287
DOI 10.1007/s10822-012-9543-4

Ligand expansion in ligand-based virtual screening using relevance feedback

Ammar Abdo · Faisal Saeed · Hentabli Hamza · Ali Ahmed · Naomie Salim

Received: 25 October 2011 / Accepted: 9 January 2012 / Published online: 17 January 2012

Abstract Query expansion is the process of reformulating an original query to improve retrieval performance in information retrieval systems. Relevance feedback is one of the most useful query modification techniques in information retrieval systems. In this paper, we introduce query expansion into ligand-based virtual screening (LBVS) using the relevance feedback technique. In this approach, a few high-ranking molecules of unknown activity are filtered from the outputs of a Bayesian inference network based on a single ligand molecule to form a set of ligand molecules. This set of ligand molecules is used to form a new ligand molecule. Simulated virtual screening experiments with the MDL Drug Data Report and maximum unbiased validation data sets show that the use of ligand expansion provides a very simple way of improving the LBVS, especially when the active molecules being sought have a high degree of structural heterogeneity. However, the effectiveness of the ligand expansion is slightly less when structurally-homogeneous sets of actives are being sought.

Keywords Virtual screening · Bayesian inference network · Ligand expansion · Nearest neighbours · Similarity searching · Drug discovery

A. Abdo (✉) · F. Saeed · H. Hamza · A. Ahmed · N. Salim
Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia, 81310 Skudai, Malaysia
e-mail: ammar_utm@yahoo.com

A. Abdo
Computer Science Department, Hodeidah University, Hodeidah,
Yemen

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]. The increasing importance of similarity searching applications is particularly due to its role in lead optimisation in drug discovery programmes, where the nearest neighbours for an initial lead compound are sought in order to find better compounds. Hence, in lead optimisation, the molecules which lie in the neighbourhood region of an active compound are selected for testing to find the one with the optimum activity. Another application of similarity searching is to predict the biological or physicochemical properties of an unknown compound, which are predicted from the properties of compounds that lie within the same neighbourhood region [2]. Underpinning these applications of molecular similarity measure is the similar property principle [3], which states that structurally-similar molecules will exhibit similar physiochemical and biological properties.

There are many studies in the literature associated with the measurement of molecular similarity [1, 4–7]. The most common approach, which we study in this paper, 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 each 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, 8].

In our attempt to improve the retrieval effectiveness of LBVS, we have introduced a Bayesian inference network (BIN) as an alternative to existing tools for LBVS [9, 10]. To this end, different models have been developed for the Bayesian network [11–14]. In our previous works, the retrieval performance of LBVS was observed to improve significantly when multiple ligand molecules were used [12, 13]. However, such information is unlikely to be available in the early stages of a drug discovery programme when just a single weak lead is available. To overcome this limitation, relevance feedback has been used [15–18].

Relevance feedback has previously been used in LBVS under the name "nearest neighbours". Turbo similarity searching (TSS) [15] and reweighted BIN [18] are two examples of using relevance feedback (nearest neighbours) information. In these approaches, a few high-ranking molecules of unknown activity were filtered from the outputs of a conventional similarity search (Tanimoto similarity/BIN) based on a single ligand molecule to form a set of ligand molecules. This set of ligand molecules is used to search the database and then the data fusion is used to combine the individual similarity searches. In addition, this set of ligand molecules is then used to reweight the original ligand molecule in BIN [18] and to enable the use of machine-learning techniques for VS without the need for an explicit training set of known active and inactive molecules [16].

In the textual information retrieval literature a standard method for improving performance is query expansion [19], which is the process of adding new terms to the original query in order to improve retrieval performance in information retrieval systems. Relevance feedback is one of the most useful query modification techniques in information retrieval systems [20, 21]. In this paper, we introduce query expansion into LBVS using the relevance feedback technique. In this approach, a few high-ranking molecules of unknown activity are filtered from the outputs of BIN based on a single ligand molecule to form a set of ligand molecules. This set of ligand molecules is then used to form a new ligand molecule. Finally, the new ligand is used by the BIN similarity system to perform an additional search (second round).

Materials and methods

To enhance the effectiveness of LBVS searches, a number of strategies are used, and these strategies can be divided into those that issue a single ligand and those that issue multiple ligands [15, 18]. In the latter case it is necessary to combine the returned ranked lists (individual similarity searches) for each ligand. In the ligand expansion method, our similarity system will automatically add certain substructural fragments (the most frequent fragments in the top-ranked molecules) to the original user ligand, which will lead to new molecules that do not match literally with the original user ligand. In the next section, we give a detailed description of how the original user ligand can be expanded using relevance feedback information to form new ligand.

Ligand expansion using relevance feedback

Using relevance feedback, a set $m \leq 100$ of high-ranking molecules of unknown activity are filtered from the outputs of the original ligand to form a set of molecules which share similar biological activity to the original ligand. A new ligand L_{avg} is then formed by taking the average of the original ligand L_O and the m set of molecule results of the original ligand according to the following formula:

$$L_{\mathrm{avg}} = \frac{1}{n_i}\left(L_o + \sum_{j=1}^{m} L_j\right)$$

where L_o is the vector of the original ligand, L_j is the vector of the jth ligand within the m, and n_i is the number of ligands containing the ith fragment. For this average, if the ith fragment does not occurred in the original ligand, we take only the average of fragments occurring in more than 70% of the total number of ligands m (the figure of 70% was selected based on experiments). The ligand expansion process is summarized in Fig. 1.

The new ligand L_{avg} is then used by the BIN similarity system to search the database one more time. The only difference between this search and the initial search (using the original ligand) is that the calculation is carried out only for those fragments common to the ligand molecule and molecule database, whereas in the initial search the calculation was conducted for the ith fragment with a value ≥ 0, given that this ith fragment occurred in the ligand molecule. However, in the ligand expansion search the calculation also conducted for those fragments occurred in less than 70% of the total number of ligands (regardless the fragment value in new ligand).

Simulated virtual screening experiments

Our first set of VS experiments used the most popular chemoinformatics database: the MDL Drug Data Report (MDDR) [22], which was employed in our previous studies of Bayesian networks [11, 14, 18], with 102,516 molecules. All molecules in the MDDR database were converted to Pipeline Pilot's ECFC4 (extended connectivity fingerprints and folded to a size of 1,024 bits) [23]. For the screening experiments, two data sets (MDDR1 and MDDR2) were chosen from the MDDR database. The MDDR1 data set

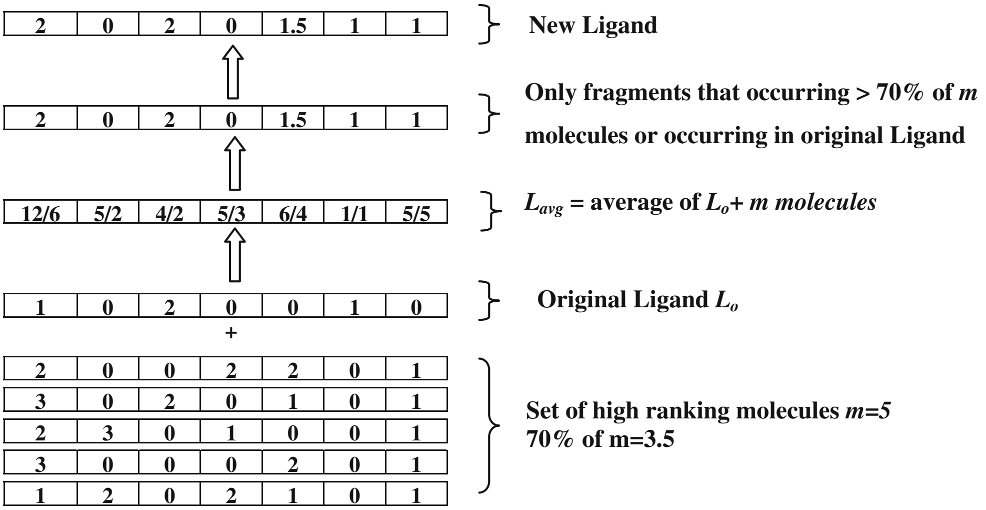

Fig. 1 Ligand expansion process

Table 1 MDDR activity classes for MDDR1 data set

Activity index	Activity class	Active molecules	Pairwise similarity (mean)
07707	Adenosine (A1) agonists	207	0.424
07708	Adenosine (A2) agonists	156	0.484
31420	Renin inhibitors	1,130	0.584
42710	Monocyclic beta-lactams	111	0.596
64100	Cephalosporins	1,346	0.512
64200	Carbacephems	113	0.503
64220	Carbapenems	1,051	0.414
64300	Penicillin	126	0.444
65000	Antibiotic, Macrolide	388	0.673
75755	Vitamin D analogous	455	0.569

Each row in the Tables 1,2, and 3 contains an activity class, the number of molecules belonging to the class, and the class's diversity, which was computed as the mean pairwise Tanimoto similarity calculated across all pairs of molecules in the class using ECFC4

Table 2 MDDR activity classes for MDDR2 data set

Activity index	Activity class	Active molecules	Pairwise similarity (mean)
09249	Muscarinic (M1) agonists	900	0.257
12455	NMDA receptor antagonists	1,400	0.311
12464	Nitric oxide synthase inhibitors	505	0.237
31281	Dopamine beta-hydroxylase inhibitors	106	0.324
43210	Aldose reductase inhibitors	957	0.370
71522	Reverse transcriptase inhibitors	700	0.311
75721	Aromatase inhibitors	636	0.318
78331	Cyclooxygenase inhibitors	636	0.382
78348	Phospholipase A2 inhibitors	617	0.291
78351	Lipoxygenase inhibitors	2,111	0.365

contains 10 homogeneous activity classes and the MDDR2 data set contains 10 heterogeneous activity classes (i.e., structurally diverse). Details of these two data sets are listed in Tables 1 and 2.

Further experiments involved the Maximum Unbiased Validation (MUV) data set (Table 3), as reported by Rohrer and Baumann [24]. This contains 17 activity classes, with each class containing up to 30 actives and 15,000 inactives. The molecules were chosen so as to ensure that VS experiments would not be affected by analogue bias or artificial enrichment, and the data set hence provides a much stiffer test of screening effectiveness than the other data sets studied here. The molecules here were again represented by ECFC4 fingerprints.

The initial screening experiments were performed using conventional BIN with 10 ligand molecules selected randomly from each activity class. Different numbers of nearest neighbours (10, 20, 50, and 100) were selected to form a set of ligand molecules for each original ligand molecule in the conventional BIN searches. These sets were used to form new ligand molecules and then initiate new searches using the conventional BIN and Tanimoto similarity system (TAN). The effectiveness of the similarity searches was evaluated using the recall, where the recall is the percentage of the actives retrieved in the top 5% of the ranked list resulting from a similarity search.

Results and discussion

The main aim of this study is to introduce the ligand expansion approach into LBVS and then identify the retrieval effectiveness of using such an approach. To achieve this aim, the ligand expansion approach was

Table 3 MUV activity classes for MUV data set

Activity index	Activity class	Active molecules	Pairwise similarity (mean)
466	S1P1 rec. (agonists)	30	0.445
548	PKA (inhibitors)	30	0.430
600	SF1 (inhibitors)	30	0.445
644	Rho-Kinase2 (inhibitors)	30	0.398
652	HIV RT-RNase (inhibitors)	30	0.416
689	Eph rec. A4 (inhibitors)	30	0.449
692	SF1 (agonists)	30	0.365
712	HSP 90 (inhibitors)	30	0.413
713	ER-a-Coact. Bind. (inhibitors)	30	0.389
733	ER-β-Coact. Bind. (inhibitors)	30	0.352
737	ER-a-Coact. Bind. (potentiators)	30	0.502
810	FAK (inhibitors)	30	0.425
832	Cathepsin G (inhibitors)	30	0.435
846	FXIa (inhibitors)	30	0.532
852	FXIIa (inhibitors)	30	0.492
858	D1 rec. (allosteric modulators)	30	0.400
859	M1 rec. (allosteric inhibitors)	30	0.386

implemented using two different similarity systems: BIN and TAN. In what follows we will refer to them as LEB and LET, respectively. The results for the searches of MDDR1, MDDR2, and MUV are shown in Tables 4, 5, and 6, respectively, using cutoffs of 5%.

Visual inspection of the recall values and the number of shaded cells in Tables 4, 5, and 6 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 [25]. This test was developed to quantify the level of agreement between multiple sets of rankings of the same set of objects; here, and in previous works [11, 14, 18], we used this approach to rank the effectiveness of different search methods. In the present context, the activity classes were considered as judges and the recall rates of the various search models as objects. The outputs of the test are the value of the Kendall coefficient, ranges from 0 (no agreement between set of ranks) to 1 (complete agreement), and the associated significance level, which indicates whether this value of the coefficient could have occurred by chance. If the value is significant (for which we used cut-off values of 0.01 or 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 MDDR1, MDDR2, and MUV) are reported in Table 7 and describe the top 5% ranking for the various search methods.

Inspection of the results reported in Table 4 shows that the conventional BIN (original ligand) produced the highest mean value compared to the ligand expansion method.

Table 4 Retrieval results of top 5% for data set MDDR1

Activity Index		BIN	Ligand Expansion-BIN				Ligand Expansion-TAN			
			10	20	50	100	10	20	50	100
07707		74.81	72.77	72.67	72.82	73.54	76.07	76.12	75.63	75.58
07708		99.61	89.03	91.23	92.97	96.90	94.52	95.35	95.68	97.35
31420		95.46	64.92	66.32	72.37	87.84	94.56	95.23	95.50	95.65
64100		92.55	85.55	86.45	88.73	90.36	90.55	91.27	91.45	91.27
64200		99.22	91.96	92.88	94.66	96.00	99.35	99.36	99.43	99.44
64220		99.20	84.38	85.80	86.96	95.27	95.45	94.82	93.66	94.02
64500		91.32	86.61	86.41	88.11	90.04	81.08	81.16	81.90	82.74
64300		94.96	64.64	66.64	72.24	81.44	75.84	77.60	82.08	80.24
65000		91.47	68.42	71.45	73.23	81.86	97.49	97.80	98.45	99.35
75755		98.33	91.63	98.13	98.15	98.39	97.95	98.13	98.15	98.15
Mean		93.69	79.99	81.80	84.02	89.16	90.29	90.68	91.19	91.38
CI	Lower	89.12	73.13	74.65	77.63	84.15	84.62	85.16	86.03	85.98
	Upper	98.27	86.85	88.95	90.41	94.18	95.95	96.21	96.36	96.78
Shaded cells		9	2	2	5	7	7	8	7	6

The left-hand part of the Tables 4, 5, and 6 contains the results for the conventional BIN, and the corresponding results when different numbers of nearest neighbours (10, 20, 50 or 100) are used to form a new ligand (ligand expansion); the right-hand part contains the corresponding results when the TAN searches are used. Each row in the Tables 4, 5, and 6 lists the recall for the top 5% of a sorted ranking when averaged over the 10 searches for each activity; 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 recall value is bold-faced; any similarity method with an average recall within 5% of the value for the best similarity method is shown lightly shaded. The bottom row in each table corresponds to the total number of shaded cells

Table 5 Retrieval results of top 5% for data set MDDR2

Activity Index	BIN	Ligand Expansion-BIN				Ligand Expansion-TAN			
		10	20	50	100	10	20	50	100
09249	27.43	31.96	33.09	34.38	33.92	27.53	27.32	26.26	25.73
12455	14.29	18.53	17.97	18.56	18.83	10.87	10.85	11.07	10.69
12464	18.13	25.58	26.13	27.46	28.17	16.37	16.87	16.03	15.65
31281	32.95	38.10	45.62	47.52	47.43	30.19	31.33	30.76	30.00
43210	15.68	14.72	15.28	15.99	17.08	16.99	17.12	16.99	16.64
71522	11.43	12.90	11.55	11.97	11.67	8.56	8.15	8.15	8.25
75721	35.01	34.57	34.55	36.52	36.20	34.38	34.16	34.28	34.14
78331	15.70	16.57	16.79	17.80	18.61	18.71	18.69	18.43	18.41
78348	20.60	18.76	20.57	21.92	22.13	27.71	27.35	27.01	26.68
78351	16.56	12.42	12.76	14.71	16.42	14.91	14.82	14.91	14.94
Mean	20.78	22.41	23.43	24.68	25.05	20.62	20.67	20.39	20.11
CI Lower	15.71	16.54	16.52	17.55	18.13	15.21	15.19	15.00	14.80
Upper	25.84	28.28	30.35	31.82	31.96	26.04	26.14	25.78	25.42
Shaded cells	2	2	2	6	8	3	3	3	3

Table 6 Retrieval results of top 5% for data set MUV

Activity Index	BIN	Ligand Expansion-BIN				Ligand Expansion-TAN			
		10	20	50	100	10	20	50	100
AC466	8.97	6.55	7.24	7.24	8.28	6.21	5.86	6.21	6.21
AC548	21.38	27.59	27.93	25.52	24.83	17.59	19.31	20.00	19.31
AC600	12.07	8.28	7.59	8.62	8.62	9.31	8.62	8.28	8.62
AC644	20.34	27.24	26.21	25.52	23.79	20.00	20.69	20.69	20.69
AC652	6.90	10.00	11.03	11.03	10.69	7.24	7.24	7.24	6.55
AC689	12.07	13.10	13.45	11.72	12.07	12.07	12.41	11.72	11.72
AC692	5.52	5.86	5.86	7.93	6.55	6.55	6.55	6.55	5.86
AC712	14.14	18.28	18.28	15.86	16.21	13.79	12.07	11.72	11.03
AC713	5.86	6.21	7.24	6.90	6.21	5.86	5.86	5.52	5.86
AC733	8.97	10.34	11.03	9.66	10.00	8.62	7.93	7.59	7.93
AC737	6.21	4.14	4.83	4.83	5.17	6.90	6.90	7.24	7.24
AC810	9.66	12.41	13.10	11.72	10.69	7.59	6.90	7.24	6.55
AC832	15.86	16.55	16.90	17.93	18.28	17.24	16.90	17.24	17.59
AC846	23.79	23.45	23.10	22.07	23.45	22.07	22.41	24.14	22.41
AC852	15.17	16.21	17.59	17.59	17.24	15.86	14.14	14.83	14.83
AC858	4.14	6.55	6.55	5.86	6.21	5.52	5.17	5.86	5.52
AC859	7.93	7.59	6.90	5.86	6.55	8.97	8.28	8.28	8.28
Mean	11.70	12.96	13.23	12.70	12.64	11.26	11.01	11.20	10.95
CI Lower	8.88	9.39	9.73	9.44	9.45	8.68	8.31	8.36	8.21
Upper	14.53	16.53	16.72	15.95	15.82	13.84	13.72	14.03	13.70
Shaded cells	3	6	11	5	5	2	1	2	2

Here, the Mean and Shaded cell figures suggest that the conventional BIN has the best overall performance at the top 5% cut-off, with the ligand expansion method performing least well. We can readily see that the ligand expansion methods (LEB100 and LET100) produced slightly lower recall rates (4.24 and 2.17%, respectively) than the conventional BIN (93.67%). The results in Tables 4 and 7 show that using the ligand expansion method to improve the retrieval effectiveness is not effective when the active molecules being sought have a high degree of structural homogeneity. However, this finding is in line with previous studies by Abdo et al. [18] and Hert et al. [15]. This is because the high degree of similarity between the molecules belonging to each class in the MDDR1 makes it very sensitive to any expansion happens to the original ligand. Especially if the new ligand contains some of the negative fragments (extracted from molecules have different activity than the original ligand). So the low performance of using such approaches is not surprised.

Table 7 Rankings of similarity approaches based on kendall W test results: MDDR1, MDDR2, and MUV top 5%

Data set	W	*p*	Ranking
MDDR1	0.564	<0.01	BIN > LET100 > LET50 > LEB100 > LET20 > LET10 > LEB50 > LEB20 > LEB10
MDDR2	0.251	<0.01	LEB100 > LEB50 > LET10 > LEB20 = BIN > LET20 > LEB10 > LET50 > LET100
MUV	0.177	<0.01	LEB20 > LEB100 > LEB10 > LEB50 > BIN > LET10 > LET50 > LET20 > LET100

The contents of the columns in table show the data set type, the value of the Kendall coefficient, the associated significance probability, and the ranking of the methods, respectively. The methods are ranked in decreasing order of screening effectiveness

The MDDR2 searches are of particular interest since they involve the most heterogeneous activity classes in the MDDR data sets, and thus provide a stiff test of the effectiveness of a screening method. Hert et al. found that TSS was not preferred to the conventional similarity search for the MDDR2 activity classes. However, when the new approach described here, ligand expansion, is used on this data set, Tables 5 and 7 show that it gives the best performance of all the methods for this data set, as shown by LEB100 and LEB50. The results in Table 5 show that ligand expansion (LEB100) is the best performing search across the 10 activity classes in terms of mean recall and number of shaded cells, with LEB50 also performing well. Table 7 shows that the value of the Kendall coefficient, 0.251, is significant at the 0.01 level of statistical significance; given that the result is significant, we can conclude that the overall ranking of the nine methods is:

$$\mathrm{LEB100} > \mathrm{LEB50} > \mathrm{LET10} > \mathrm{LEB20} = \mathrm{BIN} > \mathrm{LET20} > \mathrm{LEB10} > \mathrm{LET50} > \mathrm{LET100}.$$

If LBVS is to provide an effective tool for lead discovery, then it must be able to provide a scaffold-hopping capability for those cases where the actives belong to multiple structural classes. This was the inspiration for the design of the MDDR2 data set [16], but the MUV data set has taken this idea much further. Specifically, each of the 17 sets of 30 PubChem actives in MUV contains an average of only 1.16 molecules per scaffold, and the data set thus provides an obvious basis for further probing the effectiveness of the ligand expansion method for searching structurally diverse sets of actives. The search results for MUV are shown in Table 6, and are rather similar to those for MDDR2. A ligand expansion with BIN is again the best performing across the 17 activity classes in terms of mean recall and of shaded cells. The overall ranking of the nine different methods for the top 5% for MUV, based on the Kendall coefficient in Table 7, is:

$$\mathrm{LEB20} > \mathrm{LEB100} > \mathrm{LEB10} > \mathrm{LEB50} > \mathrm{BIN} > \mathrm{LET10} > \mathrm{LET50} > \mathrm{LET20} > \mathrm{LET100}.$$

Inspection of the results reported in Tables 4, 5, 6, and 7 shows very clearly that the ligand expansion using the relevance feedback information can significantly increase the retrieval effectiveness of VS, especially when the BIN is used for screening. Results in Tables 4, 5, 6, and 7 are presented for the original search using the conventional BIN (using an original user ligand) and the ligand expansion method (using BIN and TAN), with the 10, 20, 50 and 100 nearest neighbour molecules being used to form a new ligand molecule. When the MDDR and MUV classes and the ligand expansion with BIN are used, there is often a significant increase in the recall of the search, especially when the active molecules being sought have a high degree of structural heterogeneity (e.g., MDDR2 and MUV). However, the effectiveness of the ligand expansion method is slightly lower when structurally homogeneous sets of actives are being sought (e.g., MDDR1).

A comparison of the computed similarity values (Tables 4, 5, 6, 7, 8) for ligand expansion with BIN and TAN shows that there are often significant differences between them, especially when the active molecules being sought have a high degree of structural heterogeneity (e.g., MDDR2 and MUV). This is because the BIN treats similarity searching as an evidential reasoning process where new fragments (within the new ligand) work as new sources of evidence about ligand and molecule structure, which are combined to estimate the similarity scores. This adding of fragments contributes individually and partially to the final similarity scores, so the final similarity scores will not

Table 8 Comparison of the average percentage of active compounds retrieved using different similarity approaches: top 5%

Data set	Ligand expansion-BIN				Reweighted BIN				Group fusion			
	10	20	50	100	10	20	50	100	10	20	50	100
MDDR1	79.99	81.80	84.02	89.16	94.88	95.39	94.97	94.83	94.98	95.77	95.72	96.18
MDDR2	22.41	23.43	24.68	25.05	22.32	22.18	21.90	21.46	23.61	24.19	24.36	24.30
MUV	12.96	13.23	12.70	12.64	11.18	10.99	10.31	9.86	11.01	10.65	9.92	9.84

change in the case that some of the fragments are negative (meaning they do not belong to the activity given by the original ligand) and do not occur in the searched molecules; there will be a slight change when these fragments do occur in the searched molecules. However, TAN treats similarity searching solely as a matching process. It is certain that using such an approach with new ligand which might contain negative fragments would lead to poor results.

The consistency of different methods is also reported as 95% confidence interval. The recall rate (mean) of any method is expected to fall within this range 95% of the time. Confidence intervals help us to decide if the difference in recall rates was large enough to have practical significance. Results reported in Fig. 2 (the mean, lower and upper bounds of the confidence intervals of different methods) reveal that, we can be 95% confident that the ligand expansion method (LEB10, LEB20, LEB 50, and LEB100) perform best for MDDR2 and MUV data sets, with exception for MDDR1 data set. Therefore, on the basis of these results we can say 95% statistical certainty that the ligand expansion similarity search will do better (or at least not worse) than the conventional similarity system.

Further metrics of statistical performance analysis involved the Receiver Operating Characteristic (ROC) curve, which has been used in various fields (medicine, meteorology, etc.) [26] and also in drug discovery field [27]. A ROC curve describes the tradeoff between sensitivity and specificity, where the sensitivity is defined as the ability of the model to avoid false negatives, and the specificity relates to its ability to avoid false positives. The area under the ROC curve (AUC) is a measure of the model performance: the closer AUC is to 1, the better is the performance of the prediction. In our study we used the ROC curve to study the performance of different methods at cutoff 5%. Visual inspection of the Fig. 3 provides a preliminary indication about the quality of each method. However, the conclusion which can be drawn from Fig. 3 is the same conclusion derived from Fig. 2 and Tables 4, 5, 6, and 7.

In a recent study, Abdo et al found that using relevance feedback information and group fusion with BIN significantly outperformed all other methods in all MDDR data sets [18]. To validate the performance of ligand expansion, the new approach described here, similar experiments were repeated but using reweighted BIN and group fusion methods [18] and their results are presented in Table 8. The results reported in Table 8 show that the ligand expansion with BIN method gives the best performance of all the methods for highly structurally-diverse data sets (e.g.,

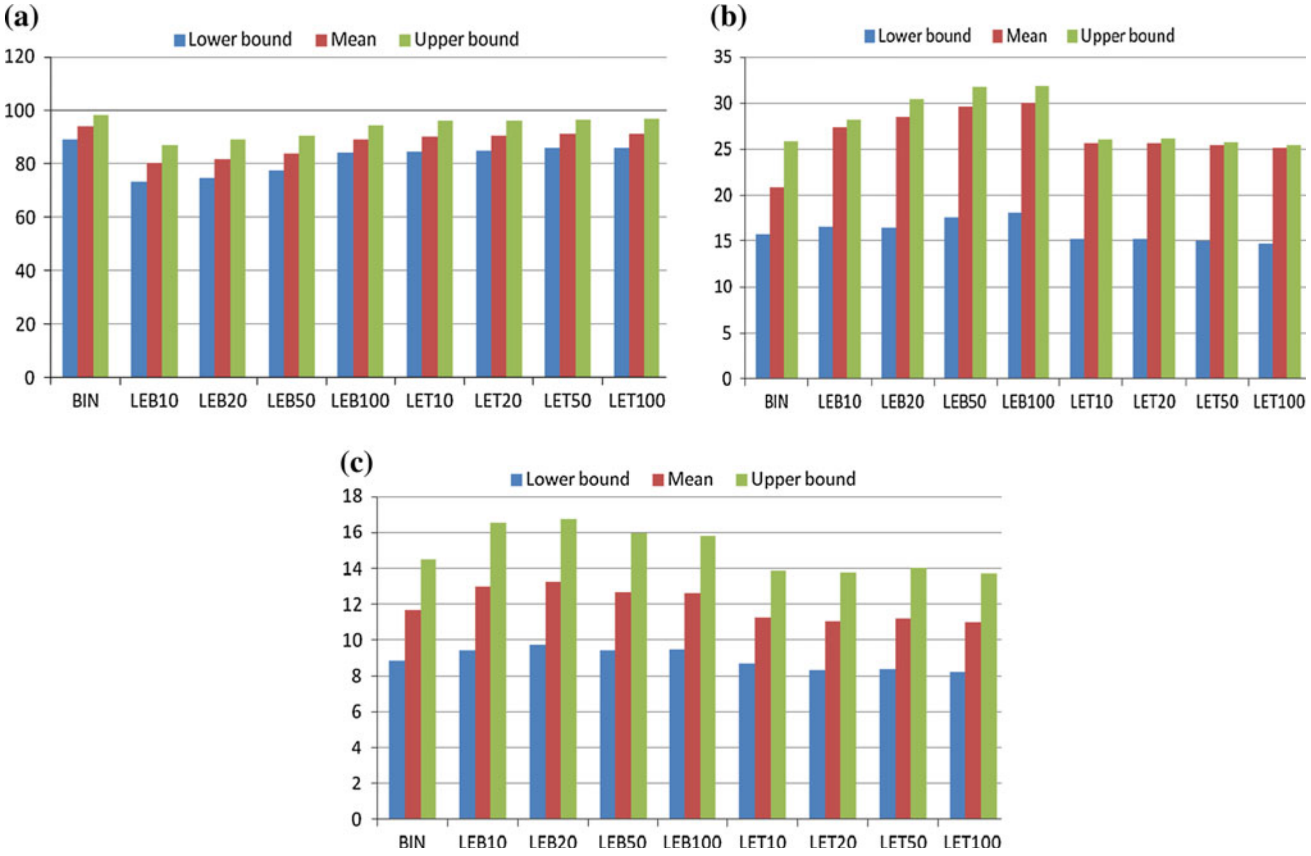

Fig. 2 Performance with 95% confidence bound for different methods with **a** MDDR1, **b** MDDR2 and **c** MUV data sets

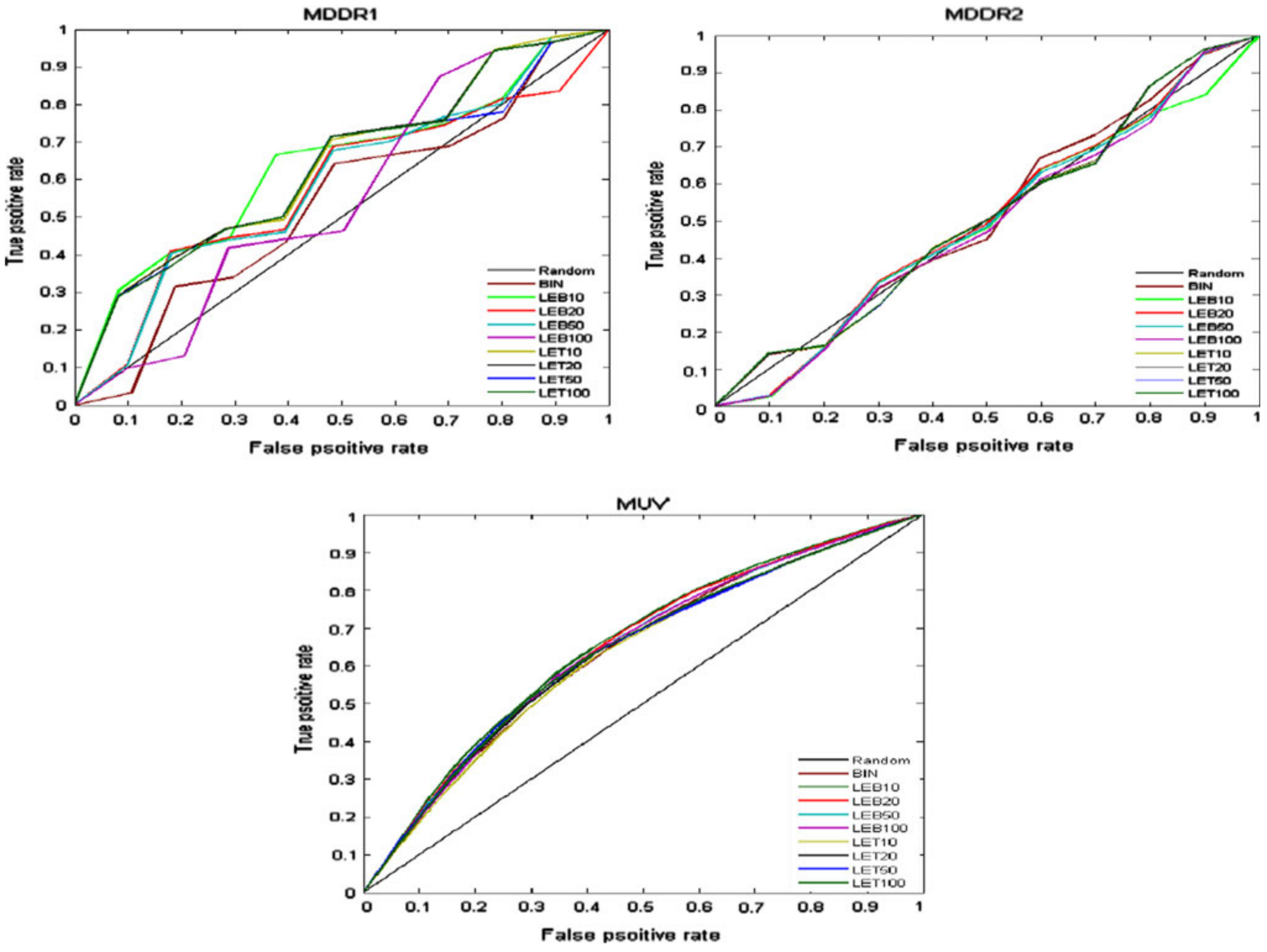

Fig. 3 ROC curves at 5% cutoff

MDDR2 and MUV). Moreover, using the ligand expansion method required only one extra search after the original search, whereas the group fusion method requires multiple search procedures (e.g., 10, 20, 50, or 100 searches). Therefore, the ligand expansion method produced better results at a lower computational cost than the use of group fusion method. Finally, a very surprising pattern of behaviour is observed in the MDDR2 and MUV results presented in Tables 5 and 6. The degrees of enhancement for these more challenging screening tasks are really remarkable. This is because this enhancement is achieved without any additional effort on the part of the user carrying out the similarity search, and at a minimal computational cost.

Conclusion

We have developed a new ligand expansion method using a Bayesian inference network. Experiments with MDDR and MUV data show that this method provides a very simple way of improving the retrieval effectiveness of LBVS, especially when the active molecules being sought have a high degree of structural heterogeneity. However, the results obtained from the experiments also show that the ligand expansion method performs least well when structurally homogeneous sets of actives are being sought.

Acknowledgments This work is supported by Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under Research University Grant Category (VOT Q.J130000.7128.00H72).

References

1. Willett P, Barnard JM, Downs GM (1998) Chemical similarity searching. J Chem Inf Comput Sci 38:983–996
2. Carhart RE, Smith DH, Venkataraghavan R (1985) Atom pairs as molecular features in structure-activity studies: definition and applications. J Chem Inf Comput Sci 25:64–73
3. Johnson MA, Maggiora GM (1990) Concepts and application of molecular similarity. Wiley, New York

4. Sheridan RP, Kearsley SK (2002) Why do we need so many chemical similarity search methods? Drug Discov Today 7:903–911
5. Nikolova N, Jaworska J (2003) Approaches to measure chemical similarity—a review. QSAR Comb Sci 22:1006–1026
6. Bender A, Glen RC (2004) Molecular similarity: a key technique in molecular informatics. Org Biomol Chem 2:3204–3218
7. Maldonado A, Doucet J, Petitjean M, Fan B-T (2006) Molecular similarity and diversity in chemoinformatics: from theory to applications. Mol Divers 10:39–79
8. Leach AR, Gillet VJ (2003) An Introduction to chemoinformatics. Kluwer, Dordrecht
9. Abdo A, Salim N (2009) Similarity-based virtual screening with a Bayesian inference network. ChemMedChem 4:210–218
10. Abdo A, Salim N (2011) Ligand-based virtual screening using Bayesian inference network. In: Library design, search methods, and applications of fragment-based drug design, vol 1076. ACS symposium series, vol 1076. American Chemical Society, pp 57–69
11. Abdo A, Salim N (2011) New fragment weighting scheme for the Bayesian inference network in ligand-based virtual screening. J Chem Inf Model 51:25–32
12. Abdo A, Salim N (2009) Bayesian inference network significantly improves the effectiveness of similarity searching using multiple 2D fingerprints and multiple reference structures. QSAR Comb Sci 28:1537–1545
13. Abdo A, Salim N (2009) Similarity-based virtual screening using Bayesian inference network: enhanced search using 2D fingerprints and multiple reference structures. QSAR Comb Sci 28: 654–663
14. Abdo A, Chen B, Mueller C, Salim N, Willett P (2010) Ligand-based virtual screening using Bayesian networks. J Chem Inf Model 50:1012–1020
15. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2005) Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information. J Med Chem 48:7049–7054
16. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2006) New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. J Chem Inf Model 46:462–470
17. Gardiner EJ, Gillet VJ, Haranczyk M, Hert J, Holliday JD, Malim N, Patel Y, Willett P (2009) Turbo similarity searching: effect of fingerprint and dataset on virtual-screening performance. Stat Anal Data Mining 2:103–114
18. Abdo A, Salim N, Ahmed A (2011) Implementing relevance feedback in ligand-based virtual screening using Bayesian inference network. J Biomol Screen 16:1081–1088
19. de Castro P, de França F, Ferreira H, Coelho G, Von Zuben F (2010) Query expansion using an immune-inspired biclustering algorithm. Nat Comput 9:579–602
20. López-Pujalte C, Guerrero-Bote VP, Moya-Anegón FD (2003) Genetic algorithms in relevance feedback: a second test and new contributions. Inf Process Manage 39:669–687
21. Taktak I, Tmar M, Hamadou A (2009) Query reformulation based on relevance feedback. In: Andreasen T, Yager R, Bulskov H, Christiansen H, Larsen H (eds) Flexible query answering systems, vol 5822. Lecture notes in computer science. Springer, Berlin, pp 134–144
22. Symyx Technologies. MDL drug data report. http://www.symyx.com/products/databases/bioactivity/mddr/index.jsp. Accessed October 20, 2011
23. Pipeline Pilot (2008) Accelrys Software Inc., San Diego
24. Rohrer SG, Baumann K (2009) Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. J Chem Inf Model 49:169–184
25. Siegel S, Castellan NJ (1988) Nonparametric statistics for the behavioral sciences. McGraw-Hill, New York
26. Swets J (1988) Measuring the accuracy of diagnostic systems. Science 240:1285–1293
27. Triballeau N, Acher F, Brabet I, Pin J-P, Bertrand H-O (2005) Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. J Med Chem 48(7):2534–2547. doi:10.1021/jm049092j