Implementing Relevance Feedback in Ligand-Based Virtual Screening Using Bayesian Inference Network

AMMAR ABDO,[1,2] NAOMIE SALIM,[1] and ALI AHMED[1]

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. The main aim of the Bayesian network model is to improve the retrieval effectiveness of similarity-based virtual screening. To this end, different models of the Bayesian network have been developed. In our previous works, the retrieval performance of the Bayesian network was observed to improve significantly when multiple reference structures or fragment weightings were used. In this article, the authors enhance the Bayesian inference network (BIN) using the relevance feedback information. In this approach, a few high-ranking structures of unknown activity were filtered from the outputs of BIN, based on a single active reference structure, to form a set of active reference structures. This set of active reference structures was used in two distinct techniques for carrying out such BIN searching: reweighting the fragments in the reference structures and group fusion techniques. Simulated virtual screening experiments with three MDL Drug Data Report data sets showed that the proposed techniques provide simple ways of enhancing the cost-effectiveness of ligand-based virtual screening searches, especially for higher diversity data sets. (*Journal of Biomolecular Screening*. 2011;16:1081-1088)

Key words: chemoinformatics, computational chemistry, structure-activity relationships, high-content screening

INTRODUCTION

OVER THE PAST FEW DECADES, drug discovery companies have used combinatorial chemistry approaches to create large and diverse libraries of structures, in which a large array of compounds are formed by combining sets of different types of reagents, called building blocks, in a systematic and repetitive way. These libraries can be used as a source of new potential drugs because compounds in the libraries can be randomly tested or screened to find a good drug compound. By increasing the capabilities of testing compounds using chemoinformatics technologies, such as high-throughput screening, it is possible to test hundreds of thousands of these compounds in a short time. Computers can be used to aid this process in a number of ways, for example, in the creation of virtual combinatorial libraries, which can be much larger than their real counterparts. These virtual libraries can be virtually screened either by docking into the active site of interest or by virtue of their similarity to a known active. Recently, searching chemical databases using a computer, instead of by experiment, has been called the virtual screening technique.[1,2]

Many virtual screening approaches have been implemented for searching chemical databases, such as the substructure search, similarity, docking, and QSAR. Similarity searching is the simplest and one of the most widely used techniques for ligand-based virtual screening in drug discovery programs.[3] There are many studies in the literature associated with the measurement of molecular similarity.[1–4] However, the most common approaches are based on two-dimensional fingerprints, with the similarity between a reference structure and a database structure computed using association coefficients such as the Tanimoto coefficient.[3]

The effectiveness of ligand-based virtual screening approaches can be enhanced by using data fusion.[5] Data fusion can be implemented using two different approaches.[5,6] The first, similarity fusion, involves searching for a single reference structure using multiple fingerprints. The similarity scores, or ranking, for each similarity measure (fingerprint) are combined to obtain the final ranking of the compounds in the database. The second approach is group fusion, in which multiple reference structures with a single similarity measure are used to search the database. Group fusion has been found to be generally more effective than similarity fusion.

In more recent studies, the Bayesian inference network (BIN) has been introduced as promising the similarity search approach.[7,8] In our previous works,[9–11] the retrieval performance of the Bayesian inference network was observed to improve significantly when multiple reference structures were used or more

[1]Information System Department, Universiti Teknologi Malaysia, Skudai, Malaysia.
[2]Computer Science Department, Hodeidah University, Hodeidah, Yemen.

Received Mar 26, 2011. Accepted for publication May 1, 2011.

Supplementary material for this article is available on the *Journal of Biomolecular Screening* Web site at http://jbx.sagepub.com/supplemental.

Journal of Biomolecular Screening 16(9); 2011

DOI: 10.1177/1087057111416658

weights were assigned to some fragments in the molecule structure. Unfortunately, such information is unlikely to be available in the early stages of a drug discovery program when just a single weak lead is available.

BIN was originally developed for text document retrieval systems.[12] Many studies in information retrieval (IR) have shown that the retrieval effectiveness of BIN can be improved by using relevance feedback technique. Relevance feedback is one of the most useful query modification techniques in IR systems.[13] It is used to improve the query formulated in the IR systems when the documents initially retrieved do not completely fulfill the user's information need. It works in the following way: A user submits a query expressing his information need to the IR system, which then ranks the documents in decreasing order of relevance to the query. The user then inspects the high-ranking documents returned by the IR system and specifies which documents are relevant and which ones are not relevant to his information need (relevance judgment). Based on the user's relevance judgment, the IR system updates the initial query, adds new terms to the query, and reweights the query terms. This process is repeated until the user is completely satisfied with the retrieved documents. Relevance feedback has been extensively studied in the literature[13,14] and is the focus of the present article.

Relevance feedback has also been used in similarity-based virtual screening under the name of its nearest neighbors. Hert et al.[15] have described an extension of similarity searching, referred to as Turbo similarity searching (TSS), in which a single active reference structure and the information about the nearest neighbors for that reference structure in a conventional similarity search are used to search the database and then data fusion is used to combine the individual similarity searches. The information about the nearest neighbors is also used to enable the use of machine-learning techniques for virtual screening without the need for an explicit training set of known active and inactive molecules.[16] The basic idea underlying TSS is the assumption that the nearest neighbors of a reference structure are actives. These assumed active structures can be used as reference structures together with the original reference structure to construct multiple reference structures that are required for group fusion. This process is performed automatically without any additional burden on the user. The nearest neighbors information has been proven to be a very effective way to enhance the effectiveness of the ligand-based virtual screening approaches.[15,16]

In this article, we enhance the screening effectiveness of the BIN using the relevance feedback information. In this approach, a few high-ranking structures of unknown activity were filtered from the outputs of BIN based on a single active reference structure to form a set of active reference structures. This set of active reference structures is used in two distinct techniques of carrying out such BIN searching: reweighting the fragments in the reference structures and group fusion methods. In the group fusion method, the BIN repeats the search using the set of reference structures and then combines the similarity scores resulting from individual reference-structure similarity searches using data fusion rules.

MATERIALS AND METHODS

This study compares the retrieval results obtained using three different similarity-based screening models. The first model was based on the BIN described by Abdo and Salim[9] that uses the Okapi (OKA1) weight, which was found to perform best in their experiments, which we shall refer to as the conventional BIN model. This model involves searching using a single reference structure. The second model was based on the BIN and relevance feedback information, which we shall refer to as the reweighted BIN model. It involves searching for a single reference structure and then using the feedback information to reweight the fragments in the reference. The third model was also based on the BIN and relevance feedback information, which we shall refer to as the group fusion model. It involves searching using multiple reference structures extracted from relevance feedback information and then combining the individual similarity searches using any fusion rules.

In what follows, we give a brief description of the BIN model on which this work is based. Full details of the BIN model are given by Abdo and Salim.[7,9] In this work, the emphasis will be on investigating the capability and performance of BIN to use the feedback information.

Conventional BIN model

Supplementary Figure S1 shows the conventional BIN model used in molecular similarity searching. It consists of three types of nodes: compound nodes as roots, fragment nodes, and a reference structure node as a 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, if multiple references are used). Each network node is binary valued, 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, we need to compute the probability in the fragment and reference nodes. Our recent work[9] showed

that one particular belief function, called OKA1,[9] was the most effective of those tested, and we have hence used this function for the experiments reported below. This function was used to compute the probability in the fragment nodes and is given by

$$bel_{OKA1}(f_i) = \alpha + (1-\alpha) \times \frac{ff_{ij}}{ff_{ij} + 0.5 + 1.5 \times \frac{|c_j|}{|C_{avg}|}} \times \frac{\log\left(\frac{m+0.5}{cf_i}\right)}{\log(m+1.0)} \times \frac{\min(ff_{ij}, ff_{ir})}{\max(ff_{ij}, ff_{ir})},$$

where α is a constant, and experiments using the Bayesian network show that the best value is 0.4.[7] ff_{ij} and ff_{ir} are the frequency of the i^{th} fragment within j^{th} compound and r reference structure, respectively; cf_i is the number of compounds containing i^{th} fragment; $|c_j|$ is the size (in terms of the number of fragments) of the j^{th} 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, specifically the SUM operator, was used. If $p_1, p_2, ..., p_n$ represent the beliefs at the fragment nodes (parent nodes of r), then the belief at r is given by

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

where n is the number of the unique fragments assigned to r reference structure, and p_i is the value of the belief function $bel(f_i)$ in i^{th} fragment node. The reader should note that in this article, we consider the use of only a single reference structure. However, the methods that we describe can be extended to multiple reference structures. This is achieved by combining the individual reference structure nodes using any of the fusion rules (e.g., MAX, SUM) as described previously for the BIN model.[10]

Reweighted BIN model

The difference between the two models (BIN and reweighted BIN) arises from the differences in the type of belief function used to produce the ranking of the compounds in the collection. In the conventional BIN model, the probability in the reference node is computed by summing up the probabilities in the fragment nodes connected to the reference node. The fragment nodes participating in the final probability are scored equally (meaning that no weight is given to any fragment node). This calculation is conducted using the SUM operator, as described above. In the reweighted BIN model, the relevance feedback information is used to assign weights to the fragment nodes associated with the reference node. Consequently, the probability value in the reference node depends on the values in the fragment nodes and the assigned weights associated to them. To illustrate how the reweighted BIN model can have different weights for fragments in the reference structure, the belief function from InQuery, the WSUM operator, has been used. If $p_1, p_2, ..., p_n$ represent the beliefs at the fragment nodes (parent nodes of r) and w_1, w_2, and w_n are the parent weights, then the belief at r is given by

$$bel_{wsum}(r) = \frac{\sum_{i=1}^{n} w_i p_i}{\sum_{i=1}^{n} w_i}.$$

To produce the relevance feedback information, we have to start with the conventional BIN model to search the database with a single active reference structure. A few high-ranking structures of unknown activity are then chosen from the outputs of conventional BIN, based on a single active reference structure, to form a set of active reference structures (relevance feedback information). The original active reference structure, together with the set of active reference structures (feedback information), form the final set that is used to reweight the BIN model. Here, the additional weights assigned to the fragments are based on the frequency of their occurrence in the set of relevance feedback structures. The weight of the i^{th} fragment w_i is given by

$$w_i = \frac{Rf_i}{MaxRf},$$

where Rf_i is the sum of the frequencies of occurrence for the fragment i^{th} in the set of relevance feedback structures, and $MaxRf$ is the maximum frequency of occurrence in the set of relevance feedback structures. Consequently, higher weight will be assigned to those fragments that occur more frequently in the relevance feedback structures.

Group fusion model

The group fusion model uses the conventional BIN model to search the database using a single active reference structure. The relevance feedback information (the nearest neighbors information) for that reference structure is also used to search the database. Data fusion rules are then used to combine the similarity scores of individual similarity searches. This model is similar to the TSS method described by Hert et al.[15] The difference is in the type of similarity-searching approach used to search the database. The basic idea underlying the reweighted BIN and group fusion model is the assumption that the relevance feedback structures (nearest neighbors) of the original reference structure are actives, as in Hert et al. These assumed active structures can be used as reference structures, together with the original reference structure, to construct the multiple

Table 1. Retrieval Results of Top 5% for Data Set DS1

		Reweighted BIN				Group Fusion			
Activity Index	**BIN**	**10**	**20**	**50**	**100**	**10**	**20**	**50**	***100***
31420	87.61	87.83	87.80	88.00	88.02	**88.57**	87.84	87.85	88.18
71523	52.72	60.44	60.51	61.62	**61.91**	54.81	57.37	59.21	57.38
37110	48.20	46.05	46.67	47.99	47.57	50.10	51.13	53.63	**53.69**
31432	77.57	83.97	85.79	88.20	**88.93**	78.34	84.19	86.73	86.82
42731	26.63	32.77	32.55	32.76	32.30	31.03	32.53	39.63	**47.44**
06233	23.49	26.67	27.16	27.60	27.50	42.52	44.19	45.98	**49.03**
06245	14.86	15.45	15.56	16.62	16.90	**20.34**	18.49	18.27	18.13
07701	27.79	31.24	31.45	31.17	31.60	30.15	31.12	31.32	**33.07**
06235	23.78	26.39	27.51	26.30	25.88	27.41	28.66	28.89	**29.83**
78374	20.20	**22.17**	21.53	20.73	20.60	21.99	21.57	20.66	19.89
78331	11.80	12.50	12.35	12.47	12.30	14.06	13.67	**15.50**	13.72
Mean	37.70	40.50	40.81	41.22	41.23	41.76	42.80	44.33	**45.20**
Shaded cells	1	3	5	3	4	3	4	6	7

reference structures that are required for the reweighted BIN and group fusion models.

Simulated virtual screening experiments

Our experiments have used the most popular chemoinformatics database: the MDL Drug Data Report (MDDR)[17] that has been used in our previous studies of Bayesian networks.[9] This database consists of 102 516 molecules. The 102 516 molecules in the database were converted to Pipeline Pilot's (Accelrys Software, San Diego, CA) ECFC4 (extended connectivity) fingerprints and folded to a size of 1024. For the screening experiments, three data sets (DS1-DS3) were chosen from the MDDR database. The data set DS1 contains 11 activity classes, with some of the classes involving actives that are structurally homogeneous and other classes involving actives that are structurally heterogeneous (i.e., structurally diverse). The DS2 data set contains 10 homogeneous activity classes, and the DS3 data set contains 10 heterogeneous activity classes. Details of these three data sets are given in **Supplementary Tables S1 to S3**. Each row of a table contains an activity class, the number of molecules belonging to the class, and the class's diversity, which is computed as the mean pairwise Tanimoto similarity, calculated across all pairs of molecules in the class using ECFP6. The pairwise similarity calculations for all data sets were conducted using Pipeline Pilot software. Full details of these data sets are given by Hert et al.[16]

Conventional BIN searches were carried out using 10 reference structures selected randomly from each activity class. Different numbers of nearest neighbors—10, 20, 50, and 100—were selected (as reference structures) for each reference structure in the conventional BIN searches. These sets were used by the reweighted BIN search and group fusion search: these searches will be referred to as BIN, RE, and GF, respectively. In the group fusion search, the outputs from sets of 10, 20, 50, or 100 searches were fused using the MAX fusion rule. The superiority of the MAX rule has been noted by previous works.[10] The effectiveness of similarity searches was evaluated by the recall, where the recall is the percentage of the actives retrieved in the top 1% or the top 5% of the ranked list resulting from a similarity search.

RESULTS AND DISCUSSION

Our purpose is to identify different approaches of using relevance feedback information in the BIN model and then identify the retrieval effectiveness of using such approaches. In this study, we tested the various BIN models (conventional BIN, reweighted BIN, and group fusion) on the MDDR database using three different data sets: DS1 to DS3.

The results of the conventional BIN, the reweighted BIN, and group fusion for the searches of DS1 to DS3 are presented in **Tables 1** to **3**, respectively, using a cutoff of 5%. The first column from the left of each table contains the results for the conventional BIN when a single reference structure is used, the second column of each table contains the corresponding results when different numbers of nearest neighbors (10, 20, 50, or 100) are used to reweight the reference fragments in the reweighted BIN model, and the last column of each table contains the corresponding results when different numbers of nearest neighbors (10, 20, 50, or 100) are used as reference structures in the group fusion model. Each row in a table lists the recall for the top 5%

Table 2. Retrieval Results of Top 5% for Data Set DS2

		Reweighted BIN				Group Fusion			
Activity Index	**BIN**	**10**	**20**	**50**	**100**	**10**	**20**	**50**	**100**
07707	74.81	75.49	75.34	75.15	74.17	**77.33**	76.8	75.53	74.47
07708	99.61	**100.00**	**100.00**	**100.00**	**100.00**	99.68	99.74	99.87	99.81
31420	95.46	95.64	95.76	96.14	96.28	96.37	96.92	97.5	**97.81**
64100	92.55	93.00	95.27	93.36	91.18	97.36	98.27	98.55	**98.73**
64200	99.22	99.29	99.41	99.52	**99.63**	99.27	99.36	99.38	99.46
64220	99.20	99.73	99.64	99.64	99.55	99.64	99.64	99.91	**99.91**
64500	91.32	92.15	93.29	93.79	91.82	96.44	98.19	**98.24**	98.12
64300	94.96	96.16	97.52	94.40	97.92	95.2	97.84	99.36	**99.52**
65000	91.47	99.07	99.38	99.48	**99.48**	90.08	92.35	89.97	94.91
75755	98.33	98.24	98.24	98.26	98.24	98.44	98.61	98.88	**99.01**
Mean	93.69	94.88	95.39	94.97	94.83	94.98	95.77	95.72	**96.18**
Shaded cells	7	8	9	8	8	9	9	9	10

Table 3. Retrieval Results of Top 5% for Data Set DS3

		Reweighted BIN				Group Fusion			
Activity Index	**BIN**	**10**	**20**	**50**	**100**	**10**	**20**	**50**	**100**
09249	27.43	30.26	31.07	31.89	32.28	30.96	32.97	38.40	**40.95**
12455	14.29	14.65	15.18	15.45	15.05	17.18	17.35	18.08	**18.62**
12464	18.13	**21.49**	21.07	19.35	17.90	19.70	18.45	17.90	17.56
31281	32.95	33.81	31.43	27.52	25.52	**43.71**	41.52	39.71	37.52
43210	15.68	15.99	15.65	16.40	16.53	18.57	**22.66**	22.13	21.04
71522	11.43	10.93	10.96	10.86	10.80	11.66	12.60	12.80	**12.90**
75721	35.01	35.37	35.50	35.64	35.64	37.09	36.90	37.24	**38.63**
78331	15.70	15.62	15.91	15.73	15.98	16.20	16.71	**17.04**	16.52
78348	20.60	**29.59**	28.93	29.25	28.62	21.59	22.34	20.70	18.96
78351	16.56	15.47	16.12	16.94	16.31	19.45	**20.39**	19.60	20.28
Mean	20.78	22.32	22.18	21.90	21.46	23.61	24.19	**24.36**	24.30
Shaded cells	0	2	2	1	1	4	6	6	6

of a sorted ranking when averaged over the 10 searches for each activity class, and the penultimate row in a table corresponds to the mean value for that similarity method when averaged over all of the activity classes for a data set. The similarity method with the best recall rate in each row is strongly shaded, and the recall value is boldfaced; 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 a table corresponds to the total number of shaded cells for each similarity method across the full set of activity classes.

Visual inspection of the recall values in **Tables 1** to **3** enables one to make comparisons between the effectiveness of the various search models. However, a more quantitative approach is possible using the Kendall *W* test of concordance.[18] This test shows whether a set of judges makes comparable judgments about the ranking of a set of objects; here, the activity classes were considered as judges and the recall rates of the various search models as objects. The output of such a test is the value of the Kendall coefficient and the associated significance level, which indicates whether this value of the coefficient could have

Table 4. Rankings of Similarity Approaches Based on Kendall *W* Test Results: DS1 to DS3 Top 1% and 5%

Data Set	*Recall Type*	*W*	*P*	*Ranking*
DS1	1%	0.264	<.01	GF50 > GF100 > GF20 > GF10 > RE50 > RE100 > RE20 > RE10 > BIN
	5%	0.352	<.01	GF100 > GF50 > GF20 > GF10 > RE50 > RE100 > RE20 > RE10 > BIN
DS2	1%	0.046	>.01	GF100 > RE50 = RE100 > GF20 > RE20 > GF50 > RE10 > BIN > GF10
	5%	0.296	<.01	GF100 > GF50 > GF20 > RE50 > RE20 > GF10 > RE100 > RE10 > BIN
DS3	1%	0.221	>.01	GF20 > GF50 > GF100 > GF10 > RE50 > RE20 > RE10 > RE100 > BIN
	5%	0.365	<.01	GF20 > GF50 > GF100 > GF10 > RE50 > RE20 > RE10 > RE100 > BIN

occurred by chance. If the value is significant (for which we used cutoff 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 DS1–DS3) are reported in **Table 4** and describe the top 1% and top 5% rankings for the various search models. In **Table 4**, the columns show the data-set type, the recall percentage, the value of the coefficient, the associated probability, and the ranking of the methods.

Some of the activity classes may contribute disproportionally to the overall value of mean recall (e.g., low-diversity activity classes). Therefore, using the mean recall value as an evaluation criterion could be impartial to some methods but not others. To avoid this bias, the effectiveness performance of different methods has been further investigated, based on the total number of shaded cells for each method across the full set of activity classes, as shown in the bottom row of **Tables 1** to **3**. These shaded cell results are listed in **Supplementary Table S4**. The results shown in the bottom rows of **Tables 1** to **3** form the lower part of results in **Supplementary Table S4**.

Inspection of the DS1 search in **Table 1** shows that the group fusion search (GF100, GF50, GF20, or GF10) produced the highest mean value compared with the BIN and reweighted BIN. In addition, according to the total number of shaded cells in **Table 1**, group fusion with 100 nearest neighbors (GF100) is the best performing search across the 11 activity classes in terms of mean recall, with GF50 also performing well. **Table 4** shows that the value of the Kendall coefficient (for DS1 top 5%), 0.352, is significant at the 0.01 level of statistical significance. Given that the result is significant, we can hence conclude that the overall ranking of the different procedures is GF100 > GF50 > GF20 > GF10 > RE50 > RE100 > RE20 > RE10 > BIN. The good performance of the group fusion method is not restricted to DS1 because it also gives the best results for the top 1% and 5% for DS2 and DS3.

The results in **Table 1** also show that the reweighted BIN and group fusion methods are always superior to the conventional BIN in their ability to identify active molecules for all activity classes, with performance being quite significant for the more heterogeneous activity classes. In only one instance (for the Renin inhibitor activity class), the increase in performance was marginal. The low performance with low-diversity activity classes can also be seen clearly through the DS2 search results reported in **Table 2**. Results in **Table 2** show that group fusion (GF100) is the best-performing search across the 10 activity classes in terms of mean recall and number of shaded cells, with reweighted BIN (RE20) also performing well. The overall ranking of the 10 different procedures based on the Kendall coefficient in **Table 4** (for DS2 top 5%) is GF100 > GF50 > GF20 > RE50 > RE20 > GF10 > RE100 > RE10 > BIN.

The DS3 searches are of particular interest because they involve the most heterogeneous activity classes in the three data sets used and thus provide a tough test of the effectiveness of a screening method. Hert et al.[16] found that TSS (group fusion) was not being preferred to the conventional similarity search for the DS3 activity classes. However, when the group fusion described here, GF, is used on this data set, **Tables 3** and **4** show that it (as shown by GF20, GF50, and GF100) gives the best performance of all the methods for this data set at both cutoffs.

Table 4 gives the level of statistical significance and the associated ranking for the set of results in **Tables 1** to **3**. It can be seen that there is a high degree of commonality in the rankings, with the group fusion method (e.g., GF100, GF50, and GF20) in particular providing a level of performance that is generally superior to the other similarity methods tested here. In addition, the results in **Supplementary Table S4** show that the higher values (i.e., corresponding to a greater number of high-effectiveness searches) tend to occur in the right-hand part of this table, corresponding to the group fusion method.

Visual inspection of the results (**Tables 1**–**4** and **Supplementary Table S4**) shows very clearly that the relevance feedback information can significantly increase the effectiveness of the BIN method. Results are presented for an original search using the conventional BIN (using a single active reference structure), the reweighted BIN, and the group fusion methods, with 10, 20, 50, and 100 nearest neighbor structures used for the reweighted BIN and group fusion methods. When the MDDR classes are used, there is often a significant increase in the recall of the search as more nearest neighbors are included in the group fusion model, with the best searches using 50 to 100 nearest neighbors. This finding is in line with a previous study by Hert

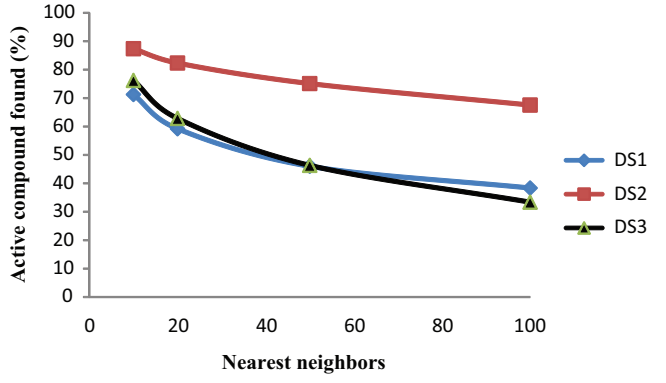

FIG. 1. Average percentage of compounds being active in different numbers of nearest neighbors.

et al.[15] Moreover, the reweighted BIN method shows superior performance compared with the conventional BIN, with little additional effort. However, both the reweighted BIN and group fusion methods show a significant enhancement because this enhancement is achieved without any additional effort on the part of the user carrying out the similarity search. A very surprising pattern of behavior is observed in the DS3 results presented in **Table 3**. The degree of enhancement for this more challenging screening task is remarkable.

In **Figure 1**, the percentage of active compounds in different data sets has been plotted as a function of its ranks (i.e., the part of the ranked list in which it would occur during a conventional BIN similarity search). This percentage was obtained by averaging the number of compounds found active at a given rank when averaged over all of the activity classes. The results are shown in **Figure 1**, which provides a better insight into why the reweighted BIN and group fusion methods work and improve the recall effectiveness. Surprisingly, the best results are generally obtained when the largest number of nearest neighbors is used, because **Figure 1** shows clearly that the percentage of active compounds decreases rapidly as one increases the number of the nearest neighbors. However, the reason the average recall does increase is that these molecules are providing useful information. This finding is in line with previous studies by Hert et al.[15] and Klon et al.[19] However, for the data sets studied here, we found that the use of 100 nearest neighbors produced better results at minimal computational cost than the use of 200 nearest neighbors, and we have hence included results only for the number 10, 20, 50, and 100 nearest neighbors.

To validate the performance of the reweighted BIN and group fusion methods, similar experiments were repeated but using the TSS[15] method, and their results are presented in **Supplementary Table S5**. Results reported **Supplementary Table S5** reveal the benefit that can be achieved using the reweighted BIN and group fusion methods. We can readily see that reweighted BIN and group fusion significantly outperform the TSS method in all data sets.

In conclusion, we have introduced two different techniques of using the relevance feedback information in ligand-based virtual screening using the BIN. The first technique is based on reference fragment reweighting, and the second is based on reference expansion (group fusion). Simulated virtual screening experiments with MDDR data sets show that the proposed techniques described here provide simple ways of enhancing the cost-effectiveness of ligand-based virtual screening in chemical databases. Our experiments also show that the increases in performance are particularly marked when the sought actives are structurally diverse. If this is not the case (low-diversity actives), then any improvement over the conventional similarity searching is likely to be marginal.

ACKNOWLEDGMENTS

This work was funded by grant 79304 from the Malaysian Ministry of Science, Technology and Innovation.

REFERENCES

1. Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening—An Overview. *Drug Discovery Today.* **1998**, *3*, 160–178.
2. Geppert, H.; Vogt, M.; Bajorath, J. R. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216.
3. Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
4. Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity—A Review. *QSAR Comb. Sci.* **2003**, *22*, 1006–1026.
5. Willett, P. Enhancing the Effectiveness of Ligand-Based Virtual Screening Using Data Fusion. *QSAR Comb. Sci.* **2006**, *25*, 1143–1152.
6. Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
7. Abdo, A.; Salim, N. Similarity-Based Virtual Screening with a Bayesian Inference Network. *Chem. Med. Chem.* **2009**, *4*, 210–218.
8. Abdo, A.; Chen, B.; Mueller, C.; Salim, N.; Willett, P. Ligand-Based Virtual Screening Using Bayesian Networks. *J. Chem. Inf. Model.* **2010**, *50*, 1012–1020.
9. Abdo, A.; Salim, N. New Fragment Weighting Scheme for the Bayesian Inference Network in Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* **2011**, *51*, 25–32.
10. Abdo, A.; Salim, N. Similarity-Based Virtual Screening Using Bayesian Inference Network: Enhanced Search Using 2D Fingerprints and Multiple Reference Structures. *QSAR Comb. Sci.* **2009**, *28*, 654–663.
11. Abdo, A.; Salim, N. Bayesian Inference Network Significantly Improves the Effectiveness of Similarity Searching Using Multiple 2D Fingerprints and Multiple Reference Structures. *QSAR Comb. Sci.* **2009**, *28*,1537–1545.

12. Turtle, H.; Croft, W. B. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.* **1991**, *9*, 187–222.
13. de Campos, L. M.; Fernandez-Luna, J. M.; Huete, J. F. Implementing Relevance Feedback in the Bayesian Network Retrieval Model. *J. Am. Soc. Inf. Sci. Technol.* **2003**, *54*, 302–313.
14. Ruthven, I.; Lalmas, M. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.* **2003**, *18*, 95–145.
15. Hert, J. R. M.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Enhancing the Effectiveness of Similarity-Based Virtual Screening Using Nearest-Neighbor Information. *J. Med. Chem.* **2005**, *48*, 7049–7054.
16. Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. 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.* **2006**, *46*, 462–470.
17. Symyx Technologies. MDL Drug Data Report. http://www.symyx.com/products/databases/bioactivity/mddr/index.jsp (accessed Feb 14, 2011).
18. Siegel, S.; Castellan, N. J. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill: New York, **1988**.
19. Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding More Needles in the Haystack: A Simple and Efficient Method for Improving High-Throughput Docking Results. *J. Med. Chem.* **2004**, *47*, 2743–2749.

Address correspondence to:
Ammar Abdo
Information System Department
Universiti Teknologi Malaysia
81310, Skudai, Malaysia

E-mail: ammar_utm@yahoo.com