New Fragment Weighting Scheme for the Bayesian Inference Network in Ligand-Based Virtual Screening

Ammar Abdo*,†,‡ and Naomie Salim†

Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310, Skudai, Malaysia and Department of Computer Science, Hodeidah University, Hodeidah, Yemen

Received June 14, 2010

Many of the conventional similarity methods assume that molecular fragments that do not relate to biological activity carry the same weight as the important ones. One possible approach to this problem is to use the Bayesian inference network (BIN), which models molecules and reference structures as probabilistic inference networks. The relationships between molecules and reference structures in the Bayesian network are encoded using a set of conditional probability distributions, which can be estimated by the fragment weighting function, a function of the frequencies of the fragments in the molecule or the reference structure as well as throughout the collection. The weighting function combines one or more fragment weighting schemes. In this paper, we have investigated five different weighting functions and present a new fragment weighting scheme. Later on, these functions were modified to combine the new weighting scheme. Simulated virtual screening experiments with the MDL Drug Data Report[23] and maximum unbiased validation data sets show that the use of new weighting scheme can provide significantly more effective screening when compared with the use of current weighting schemes.

INTRODUCTION

Similarity methods may be the simplest tools for ligand-based virtual screening, the most widely used of which is similarity searching based on two-dimensional (2D) fingerprints, since it has been shown to offer a high level of screening effectiveness and is computationally feasible.[1−5] This approach uses molecules characterized by 2D fingerprints that encode the presence of 2D fragment substructures in a molecule. The similarity between two molecules is then computed using the number of substructural fragments common to a pair of structures and a simple association coefficient, most usually the Tanimoto coefficient (TAN).[6]

Many of the conventional similarity methods assume that molecular fragments that do not relate to the biological activity carry the same weight as the important ones. It is typical for the chemist to regard some features/fragments as being more important than others through the chemist structure diagrams, such as functional groups. Thus, giving more weight to those fragments is more important than giving to others. As a result, a match between a pair of molecules on a highly weighted fragment would make a greater contribution to the overall similarity than a match on a less important fragment.

In this paper, we focus on the Bayesian inference network (BIN) model, which models molecules and reference structures as probabilistic inference networks. BINs were originally developed for text document retrieval and have become popular in the information retrieval field.[7,8] Recently, BIN was introduced to the chemoinfomatics area as a promising similarity search approach. The first work which introduced the BIN for molecular similarity searching was conducted by Abdo and Salim.[9−11] This work showed that using the BIN substantially outperformed the conventional similarity approaches. This work has been repeated later by Chen et al.[12] with different databases. These two studies showed that BIN substantially outperforms the TAN method, especially when the active molecules being sought have a high degree of structural homogeneity but have been found to perform less well with structurally heterogeneous sets of actives. To overcome this limitation, an alternative Bayesian network model called Bayesian belief network (BBN) has been introduced by Abdo et al.[13] In their experiments, BBN was compared with BIN and TAN using the MDL Drug Data Report (MDDR), WOMBAT, and maximum unbiased validation (MUV) data sets, and they have repeated similar experiments when the raw fragment frequencies in the BBN, BIN, and TAN are replaced by square roots of those frequencies. They found that the performance of BBN and BIN approaches was not obviously superior to the TAN approach that used the square root of fragment occurrence frequencies. However, the performance of the BIN model depends on the fragment weighting function that is used to differentiate between different fragments in a molecule, based on how important they are in determining the similarity of that molecule with another molecule. Certain molecular fragments can be emphasized by associating higher weights to them when calculating similarity. Usually, the weighting functions consist of one or more fragment weighting schemes.[10,14]

In the chemoinformatics field, many of the weighting functions originate from the information retrieval field, where many similarities have been identified between them.[15] These analogies have provided the basis for the work in this paper, which is to introduce a new fragment weighting scheme and

* Corresponding author. E-mail: Ammar_utm@yahoo.com. Telephone: 006-017-7425041.
† Universiti Teknologi Malaysia.
‡ Hodeidah University.

10.1021/ci100232h
Published on Web 12/14/2010

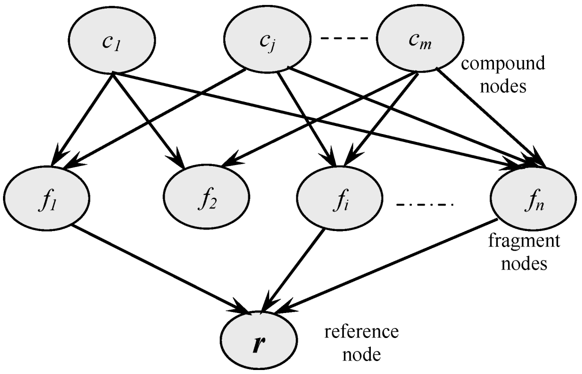

Figure 1. Bayesian inference network model.

then develop new weighting functions for a molecular similarity method based on the Bayesian network. In what follows, we outline the BIN model and investigate five different weighting functions with two popular fragment weighting schemes.

MATERIALS AND METHODS

In what follows, we give a brief description of the Bayesian network model on which this work is based (see refs 9−12, 14, and 16 for more details). In this paper, the emphasis will be on investigating the performance of the BIN model using different fragment weighting functions.

Bayesian Inference Network Model. Figure 1 shows the Bayesian network model used in molecular similarity searching. It consists of three types of nodes: compound nodes (c_j's) as roots, fragment nodes (f_i's), and a reference structure node (r) as leaf (where 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 compound node has a prior probability associated with it that describes the probability of observing that compound. Each fragment node has one or more compound nodes as parents and one reference structure node as child (or more in case multiple references used). Each network node is binary valued, taking one of the two values from the set {true, false}. The probability that the reference is satisfied given a particular compound c_j is computed by setting the value of the c_j to true and computing the probabilities associated with each fragment node connected to the reference structure node. Probabilities are computed using the fragment weighting functions.

Fragment Weighting Functions. We need to provide estimates that characterize the dependence of the random variables (nonroot nodes) in our model. The dependency can be quantified by the conditional probability, whereas the conditional probability can be estimated by many types of fragment weighting functions. Proper fragment weighting can greatly improve the performance of the BIN method. A weighting function is composed of two different types of fragment weighting schemes: local and global fragment weights. The local weight is computed according to the frequencies of fragments in the given compound or the reference structure, whereas the global weight is based on the compound collection regardless of whether we are weighting compounds or reference structures. In our model, estimates are required for the nonroot nodes: fragment and reference structure nodes.

Fragment Nodes. To estimate the probability that a fragment node is good for discriminating between different compounds, different weighting schemes can be incorporated in the weighting function. In this study, we have investigated five different weighting functions as in eqs 1−5. These have all been used previously in information retrieval to model information about the occurrences of textual keywords (terms) but have been modified here to model information about the occurrences of substructural fragments. These weighting functions are given by

$$\mathrm{bel}_{\mathrm{STD}}(f_i) = \alpha + (1 - \alpha) \times \frac{ff_{ij}}{\max ff_j} \times \frac{\log\left(\dfrac{m + 0.5}{cf_i}\right)}{\log(m + 1.0)} \tag{1}$$

$$\mathrm{bel}_{\mathrm{OKA}}(f_i) = \alpha + (1 - \alpha) \times \frac{ff_{ij}}{ff_{ij} + 0.5 + 1.5 \times \dfrac{|c_j|}{|C_{\mathrm{avg}}|}} \times \frac{\log\left(\dfrac{m + 0.5}{cf_i}\right)}{\log(m + 1.0)} \tag{2}$$

$$\mathrm{bel}_{\mathrm{AUG}}(f_i) = \alpha + (1 - \alpha) \times \left(0.5 + 0.5 \times \frac{ff_{ij}}{\max ff_j}\right) \times \frac{\log\left(\dfrac{m + 0.5}{cf_i}\right)}{\log(m + 1.0)} \tag{3}$$

$$\mathrm{bel}_{\mathrm{SQR}}(f_i) = \alpha + (1 - \alpha) \times (\sqrt{ff_{ij} - 0.5} + 1) \times \frac{\log\left(\dfrac{m + 0.5}{cf_i}\right)}{\log(m + 1.0)} \tag{4}$$

$$\mathrm{bel}_{\mathrm{SMO}}(f_i) = \lambda \times \frac{ff_{ij}}{|c_j|} + (1 - \lambda) \times \frac{CF_i}{|C|} \tag{5}$$

where α and λ are constants and experiments using the Bayesian network show that the best values are 0.4 and 0.6, respectively,[10−12] ff_{ij} is the frequency of the i^{th} fragment within j^{th} compound, cf_i is the number of compounds containing i^{th} fragment, CF_i is the sum of the frequencies of occurrence for the fragment i^{th} in the database, $\max ff_j$ is the maximum frequency of occurrence in j^{th} compound, $|c_j|$ is the size (in terms of number of fragments) of the j^{th} compound, $|C_{\mathrm{avg}}|$ is the average size of all the compounds in the database, $|C|$ is the size (in terms of number of fragments) of all the compounds in the database, and m is the total number of compounds.

The above weighting functions are composed of two different types of fragment weighting: local and global. The local weight of any fragment (in the given compound or the reference structure fingerprints) is given by a function of how many times this fragment occurs in a compound/reference structure, whereas the global weight is given by a function of how many times this fragment occurs in the entire compound collection. Thus, the first part in the above functions represents the local fragment weight, whereas the second part represents the global fragment weight. Readily we can see that the preliminary part of local weight functions is ff_{ij}, while cf_i is the preliminary part of global weight functions.

The first weighting function above was originally used in the InQuery retrieval system.[8,17] The second weighting function is closely related to the first function in eq 1 but has been adapted from the equation developed and used in the OKAPI retrieval system.[18,19] An alternative, and also closely related, function in eqs 3 and 4 has been used in text retrieval experiments.[20,21] Finally, Metzler and Croft have used a weighting function, called smoothing function, from studies of language modeling, which is a formal probabilistic framework for studies in speech recognition and statistical machine translation.[22] However, more details about these functions are provided in the cited literature, and we have used all five of them in our experiments; they will be referred to as STD (for standard),[8,17] OKA (for OKAPI),[18,19] AUG (for augmented normalized term frequency),[20] SQR (for square root),[21] and SMO (for smoothing function),[22] respectively.

There are clearly only two different weighting schemes used within the above functions, fragment frequency (local weight) and inverse fragment frequency (global weight). In addition to these two weighting schemes, we have introduced a new weighting scheme. This scheme is given by

$$\frac{\min(ff_{ij},ff_{ir})}{\max(ff_{ij},ff_{ir})} \tag{6}$$

Here, the ff_{ir} is the frequencies of the i^{th} fragment within r reference structure. The value of this scheme is "1" in case the i^{th} fragment frequency in compound and reference structure is equal; otherwise the value will be less than "1". Fortunately, this scheme captures the differences between the fragment frequencies in compound and reference structures, whereas other schemes do not. The various weighting functions have been modified to include the new weighting scheme. In what follows we will refer to them as STD1, OKA1, AUG1, SQR1, and SMO1, respectively. These modified weighting functions are given by

$$\mathrm{bel}_{\mathrm{STD1}}(f_i) = \alpha + (1-\alpha) \times \frac{ff_{ij}}{\max ff_j} \times \frac{\log\left(\dfrac{m+0.5}{cf_i}\right)}{\log(m+1.0)} \times \frac{\min(ff_{ij},ff_{ir})}{\max(ff_{ij},ff_{ir})} \tag{7}$$

$$\mathrm{bel}_{\mathrm{OKA1}}(f_i) = \alpha + (1-\alpha) \times \frac{ff_{ij}}{ff_{ij} + 0.5 + 1.5 \times \dfrac{|c_j|}{|C_{\mathrm{avg}}|}} \times \frac{\log\left(\dfrac{m+0.5}{cf_i}\right)}{\log(m+1.0)} \times \frac{\min(ff_{ij},ff_{ir})}{\max(ff_{ij},ff_{ir})} \tag{8}$$

$$\mathrm{bel}_{\mathrm{AUG1}}(f_i) = \alpha + (1-\alpha) \times \left(0.5 + 0.5 \times \frac{ff_{ij}}{\max ff_j}\right) \times \frac{\log\left(\dfrac{m+0.5}{cf_i}\right)}{\log(m+1.0)} \times \frac{\min(ff_{ij},ff_{ir})}{\max(ff_{ij},ff_{ir})} \tag{9}$$

$$\mathrm{bel}_{\mathrm{SQR1}}(f_i) = \alpha + (1-\alpha) \times (\sqrt{ff_{ij} - 0.5} + 1) \times \frac{\log\left(\dfrac{m+0.5}{cf_i}\right)}{\log(m+1.0)} \times \frac{\min(ff_{ij},ff_{ir})}{\max(ff_{ij},ff_{ir})} \tag{10}$$

$$\mathrm{bel}_{\mathrm{SMO1}}(f_i) = \left(\lambda \times \frac{ff_{ij}}{|c_j|} + (1-\lambda) \times \frac{CF_i}{|C|}\right) \times \frac{\min(ff_{ij},ff_{ir})}{\max(ff_{ij},ff_{ir})} \tag{11}$$

Reference Node. We need to encode the dependency of the reference structure formulation upon the fragment nodes. To encode this dependency, we have used the belief function bel(r) from InQuery, specifically the SUM operator. 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

$$\mathrm{bel}_{\mathrm{sum}}(r) = \frac{\sum_{i=1}^{n} p_i}{n} \tag{12}$$

where n is the number of the unique fragments assigned to r reference structure, and p_i is the value of any belief functions bel(f_i) in i^{th} fragment node. The reader should note that in this paper, 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 the weighted max (WMAX) or weighted sum (WSUM) operators, as described previously for the BIN model.[11]

EXPERIMENTAL SECTION

Our experiments have used the most popular chemoinformatics database: the MDDR[23] that has been used in our previous studies of Bayesian networks.[12,13] This database consisted of 102 516 compounds, with all compounds in the database converted to Pipeline Pilot's ECFC4 (extended connectivity) fingerprints and folded to a size of 1024.[24] For the screening experiments, three data sets (DS1−DS3) were chosen (as described by Hert et al.)[25] from the MDDR database. The data set DS1 contains 11 activity classes, with some of the classes involving actives that are structurally homogeneous and with others 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. However, there are a few differences between DS2 and DS3 data sets used here and those used by Hert et al.[25] These differences are devoted to two different activity classes in the DS2 data set and the size of activity classes in DS2 and DS3.

Further experiments involved the maximum unbiased validation (MUV) data set reported recently by Rohrer and Baumann.[26] This contains 17 activity classes, with each class containing 30 actives and 15 000 inactives. The molecules have been chosen to ensure that virtual screening experiments will not be affected by analogue bias or artificial enrichment, and hence, the data set provides a much stiffer test of screening effectiveness than the other data sets studied here. The molecules here were again represented by ECFC4

Table 1. MDDR Activity Classes for DS1

activity index	activity class	active molecules	pairwise 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	5HT3 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

Table 2. MDDR Activity Classes for DS2

activity index	activity class	active molecules	pairwise similarity (mean)
07707	adenosine (A1) agonists	207	0.229
07708	adenosine (A2) agonists	156	0.305
31420	renin inhibitors	1130	0.290
64100	monocyclic β-lactams	111	0.361
64200	cephalosporins	1346	0.336
64220	carbacephems	113	0.322
64500	carbapenems	1051	0.269
64300	penicillin	126	0.260
65000	antibiotic, macrolide	388	0.305
75755	vitamin D analogous	455	0.386

Table 3. MDDR Activity Classes for DS3

activity index	activity class	active molecules	pairwise similarity (mean)
09249	muscarinic (M1) agonists	900	0.111
12455	NMDA receptor antagonists	1400	0.098
12464	nitric oxide synthase inhibitors	505	0.102
31281	dopamine β-hydroxylase inhibitors	106	0.125
43210	aldose reductase inhibitors	957	0.119
71522	reverse transcriptase inhibitors	700	0.103
75721	aromatase inhibitors	636	0.110
78331	cyclooxygenase inhibitors	636	0.108
78348	phospholipase A2 inhibitors	617	0.123
78351	lipoxygenase inhibitors	2111	0.113

fingerprints. Details of these four data sets are given in Tables 1−4. Each row of a table 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 ECFP6. The pairwise similarity calculations for all data sets were conducted using Pipeline Pilot software.[24] The screening experiments were performed with 20 reference structures selected randomly from each activity class. The recall results were averaged over each such set of active compounds, where the recall is the percentage of the actives retrieved in the top 1% or 5% of the ranked list resulting from a similarity search.

RESULTS AND DISCUSSION

Table 4. MUV Activity Classes for DS4

activity index	activity class	active molecules	pairwise similarity (mean)
466	S1P1 rec. (agonists)	30	0.117
548	PKA (inhibitors)	30	0.128
600	SF1 (inhibitors)	30	0.123
644	ρ-kinase2 (inhibitors)	30	0.122
652	HIV RT-RNase (inhibitors)	30	0.099
689	Eph rec. A4 (inhibitors)	30	0.113
692	SF1 (agonists)	30	0.114
712	HSP 90 (inhibitors)	30	0.106
713	ER-a-Coact. Bind. (inhibitors)	30	0.113
733	ER-β-Coact. Bind. (inhibitors)	30	0.114
737	ER-a-Coact. Bind. (potentiators)	30	0.129
810	FAK (inhibitors)	30	0.107
832	cathepsin G (inhibitors)	30	0.151
846	FXIa (inhibitors)	30	0.161
852	FXIIa (inhibitors)	30	0.150
858	D1 rec. (allosteric Modulators)	30	0.111
859	M1 rec. (allosteric inhibitors)	30	0.126

Our purpose is to identify the capability of the BIN method to employ different weighting functions (using new and current weighting schemes) and then identify the retrieval effectiveness of using such functions. In this study, we tested the various weighting functions (modified and unmodified functions) on the MDDR and MUV databases using four different data sets DS1−DS4. In order to validate the performance of BIN with the new weighting functions (functions with a new scheme), similar experiments were repeated using various weighting functions, giving a total of 10 possible weighting functions.

The results for the searches of DS1−DS4 are shown in Tables 5−8, respectively, using a cutoff of 1%. The left-hand side of each table contains the results for the unmodified weighting functions (STD, OKA, AUG, SQR, and SMO), the right-hand side contains the corresponding results when the modified weighting functions are used (STD1, OKA1, AUG1, SQR1, and SMO1). Each row in a table lists the recall for the top 1% of a sorted ranking when averaged over 20 reference structures for each activity class. The penultimate row in a table corresponds to the mean value for that weighting function when averaged over all of the activity classes for a data set. The weighting function with the best recall rate in each row is strongly shaded, and the recall value is bold faced; any weighting function with an average recall within 5% of the value for the best weighting function is shown slightly shaded. The bottom row in a table corresponds to the total number of shaded cells for each weighting function across the full set of activity classes.

To determine which of the weighting functions performs best, a Kendall *W* test of concordance was applied to the retrieval rates.[27] 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 as judges and the recall rates of the various weighting functions 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 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−DS4) are reported in Table 9 and describe the top 1% and 5% rankings for the various weighting functions. In Table 9, the columns show the data set type, the recall percentage, the value of

Table 5. Retrieval Results of Top 1% for DS1

Activity Index	STD	OKA	AUG	SQR	SMO	STD1	OKA1	AUG1	SQR1	SMO1
31420	52.88	72.31	52.88	39.39	38.90	63.74	75.97	**76.14**	75.84	53.73
71523	9.05	22.40	9.05	14.15	18.83	13.09	**25.19**	24.01	24.73	19.67
37110	14.76	20.32	14.76	11.20	9.70	18.44	**23.39**	21.68	20.69	11.50
31432	27.32	**41.39**	27.31	37.27	31.05	30.19	40.38	40.86	41.34	36.00
42731	9.09	16.63	9.09	13.84	10.76	12.46	**17.58**	16.68	17.11	13.78
06233	14.97	14.39	14.97	10.47	10.17	**16.52**	15.87	14.37	13.48	15.37
06245	6.87	9.57	6.86	7.26	6.42	7.11	**9.58**	9.32	9.47	8.52
07701	7.86	10.69	7.86	7.61	6.56	8.36	**11.29**	11.17	11.07	8.52
06235	10.61	**11.96**	10.61	8.69	9.01	10.79	11.68	10.67	10.48	10.35
78374	8.43	13.89	8.43	10.80	14.21	10.27	16.11	12.42	13.81	**17.27**
78331	4.39	6.49	4.39	4.44	6.87	4.85	6.49	5.44	5.31	**8.35**
Mean	15.11	21.82	15.11	15.01	14.77	17.80	**23.05**	22.07	22.12	18.46
Shaded cells	0	2	0	0	0	1	9	5	6	2

Table 6. Retrieval Results of Top 1% for DS2

Activity Index	STD	OKA	AUG	SQR	SMO	STD1	OKA1	AUG1	SQR1	SMO1
07707	61.55	62.84	61.55	59.98	54.37	62.16	63.25	**63.45**	63.20	57.45
07708	73.81	96.19	73.81	89.87	56.00	78.26	96.23	98.16	**99.16**	56.87
31420	56.49	71.50	56.48	39.43	41.21	68.25	**78.43**	77.23	76.11	57.19
64100	64.91	83.82	64.91	82.45	**88.27**	67.23	82.36	80.05	81.50	86.32
64200	77.42	85.97	77.42	82.27	61.22	78.36	85.94	**88.45**	88.15	69.40
64220	63.48	67.50	63.48	56.38	51.83	64.91	**68.44**	64.69	64.06	60.36
64500	61.62	68.00	61.62	48.92	40.89	64.59	**69.15**	67.47	66.89	47.66
64300	72.36	70.52	72.32	56.32	34.64	**73.80**	70.20	70.20	67.28	44.16
65000	55.22	85.50	55.22	81.30	73.00	67.14	85.59	84.57	**85.97**	80.25
75755	96.73	97.75	96.73	96.45	94.75	97.26	**98.02**	97.62	97.29	95.25
Mean	68.36	78.96	68.35	69.34	59.62	72.20	**79.76**	79.19	78.96	65.49
Shaded cells	3	8	3	1	2	3	9	8	7	2

Table 7. Retrieval Results of Top 1% for DS3

Activity Index	STD	OKA	AUG	SQR	SMO	STD1	OKA1	AUG1	SQR1	SMO1
09249	13.56	12.31	13.56	7.24	10.11	15.10	13.28	11.04	9.65	**15.48**
12455	6.27	8.90	6.27	4.51	7.68	7.60	9.55	7.76	7.07	**9.90**
12464	8.54	7.87	8.54	3.12	5.95	**10.23**	8.32	6.75	6.07	8.43
31281	17.62	19.24	17.62	11.38	19.48	19.86	20.67	16.29	15.43	**27.05**
43210	4.49	7.71	4.49	4.58	6.30	5.48	7.82	6.94	6.63	**8.40**
71522	5.92	5.46	5.92	3.41	3.08	**6.40**	5.49	5.18	4.69	4.47
75721	14.61	18.84	14.61	12.84	18.19	15.84	18.76	16.69	15.91	**19.00**
78331	5.35	7.59	5.35	5.41	8.75	6.06	7.53	6.42	6.37	**9.31**
78348	5.48	8.43	5.48	7.70	**12.23**	5.47	8.43	6.66	7.30	10.51
78351	6.35	10.00	6.35	5.40	9.32	8.28	11.68	9.57	9.04	**12.49**
Mean	8.82	10.64	8.82	6.56	10.11	10.03	11.15	9.33	8.82	**12.50**
Shaded cells	0	1	0	0	2	3	1	0	0	7

the coefficient, the associated probability, and the ranking of the functions.

Since some of the activity classes may contribute disproportionally to the overall value of mean recall (e.g., low diverse activity classes). Therefore, using the mean recall value as evaluation criterion could be impartial to some weighting functions but not others. To avoid this bias, the effectiveness performance of different weighting functions has been further investigated based on the total number of shaded cells for each weighting functions across the full set of activity classes, as shown in the bottom row of Tables 5−8. These shaded cell results are listed in Table 10 (e.g., the results shown in the bottom rows of Tables 5−8 form the upper part of results in Table 10).

Inspection of the results reported in Table 5 shows that the new weighting functions (STD1, OKA1, AUG1, SQR1, and SMO1) produced the highest mean value compared to the other weighting functions. In addition, according to the total number of shaded cells in Table 5, OKA1 is the best performing weighting function across the 11 activity classes in terms of mean recall, with AUG1 and SQR1 also performing well. Table 9 shows that the value of the Kendall coefficient (for DS1 top 1%), 0.554, is significant at the 0.001 level of statistical significance; given that the result is

Table 8. Retrieval Results of Top 1% for DS4

Activity Index	STD	OKA	AUG	SQR	SMO	STD1	OKA1	AUG1	SQR1	SMO1
466	2.41	3.62	2.41	3.28	2.41	2.59	**3.62**	3.10	3.28	2.59
548	9.83	11.55	9.83	10.17	8.97	10.34	**11.72**	10.86	10.86	10.52
600	2.59	2.59	2.59	2.76	1.90	**3.10**	2.93	**3.10**	**3.10**	2.41
644	4.48	7.59	4.48	7.24	8.45	5.69	8.28	7.41	7.93	**8.79**
652	3.10	**3.97**	3.10	**3.97**	1.72	2.41	2.93	2.93	3.45	2.24
689	4.31	4.66	4.31	**5.52**	4.31	3.62	4.66	4.48	5.00	2.07
692	**1.90**	**1.90**	**1.90**	1.72	1.38	1.72	1.55	1.72	1.55	1.38
712	2.07	5.17	2.07	5.17	**5.34**	1.90	4.66	3.62	4.14	4.48
713	2.07	**2.41**	2.07	**2.41**	2.24	1.72	2.07	2.07	2.07	1.72
733	3.10	3.45	3.10	**3.62**	2.93	3.10	3.28	3.10	3.28	2.07
737	1.03	1.72	1.03	2.07	1.55	1.03	1.90	2.24	**2.59**	1.38
810	3.45	3.62	3.45	2.59	2.59	3.10	3.62	3.62	**4.31**	2.41
832	6.72	9.14	6.72	8.28	7.76	7.41	**9.31**	9.14	8.97	8.97
846	5.86	10.52	5.86	9.66	7.76	7.07	**11.38**	11.21	**11.38**	6.90
852	6.90	8.10	6.90	8.10	**8.79**	7.59	8.28	8.10	8.28	8.28
858	2.24	1.72	2.24	1.72	2.24	2.24	1.90	1.90	1.90	**2.76**
859	1.55	1.38	1.55	1.38	1.21	**1.90**	1.03	1.72	1.55	1.21
Mean	3.74	4.89	3.74	4.69	4.21	3.91	4.89	4.72	**4.92**	4.13
Shaded cells	1	8	1	5	3	2	4	3	5	3

Table 9. Rankings of Weighting Functions Based on Kendall *W* Test Results for DS1−DS4 Top 1% and 5%

data set	recall type	*W*	*p*	ranking
DS1	1%	0.554	<0.001	OKA1 > OKA > AUG1 > SQR1 > SMO1 > STD1 > SQR > SMO > STD > AUG
	5%	0.383	<0.001	OKA1 > OKA > SQR1 > SMO1 > AUG1 > STD1 > SQR > STD > SMO > AUG
DS2	1%	0.544	<0.001	OKA1 > OKA > AUG1 > SQR1 > STD1 > STD > SQR > AUG > SMO1 > SMO
	5%	0.661	<0.001	OKA > AUG1 > SQR1 > OKA1 > STD1 > SQR > STD = AUG > SMO = SMO1
DS3	1%	0.449	<0.001	SMO1 > OKA1 > OKA > STD1 > SMO > AUG1 > STD = AUG > SQR1 > SQR
	5%	0.454	<0.001	SMO1 > OKA1 > SMO = OKA > STD1 > STD > AUG1 > AUG > SQR1 > SQR
DS4	1%	0.231	<0.001	SQR1 > OKA > OKA1 > AUG1 > SQR > SMO > STD1 > SMO1 > STD = AUG
	5%	0.232	<0.001	OKA > OKA1 > SQR > SQR1 > SMO > SMO1 > AUG1 > STD1 > STD = AUG

Table 10. Numbers of Shaded Cells for Mean Recall of Actives Using Different Weighting Functions for DS1−DS4 Top 1% and 5%

data set	STD	OKA	AUG	SQR	SMO	STD1	OKA1	AUG1	SQR1	SMO1
					Top 1%					
DS1	0	2	0	0	0	1	9	5	6	2
DS2	3	8	3	1	2	3	9	8	7	2
DS3	0	1	0	0	2	3	1	0	0	7
DS4	1	8	1	5	3	2	4	3	5	3
					Top 5%					
DS1	0	6	0	2	0	1	7	3	5	5
DS2	7	10	7	6	2	7	9	10	10	2
DS3	0	1	0	0	2	2	1	0	0	7
DS4	0	6	0	7	8	1	5	3	6	3

significant, we can hence conclude that the overall ranking of the 10 functions is OKA1 > OKA > AUG1 > SQR1 > SMO1 > STD1 > SQR > SMO > STD > AUG. The good performance for OKA1 is not restricted to DS1 since it also gives the best results for the top 1% for DS2.

Results in Table 6 show that OKA1 is the best performing weighting function across the 10 activity classes in terms of mean recall and number of shaded cells, with AUG1 and SQR1 also performing well. The overall ranking of the 10 functions based on the Kendall coefficient in Table 9 (for DS2 top 1%) is OKA1 > OKA > AUG1 > SQR1 > STD1 > STD > SQR > AUG > SMO1 > SMO.

The DS3 searches are of particular interest since they involve the most heterogeneous activity classes in the first three data sets used and thus provide a stiff test of the effectiveness of new weighting functions. Results in Table 7 show that SMO1 is the best performing weighting function across the 10 activity classes for this data set. The superiority of the SMO1 weighting function for searching a highly diverse data set (DS3) is in line with a previous study by Chen et al.[12]

To prove that the new weighting functions provide an effective tool for lead discovery, the ability of different weighting functions to identify structurally diverse different molecules that represent different lead series but share similar biological activity (a scaffold-hooping) has been investigated. This had been the inspiration for the design of the DS3 data set, but the DS4 (MUV) data set has taken this idea much further. Specifically, each of the 17 sets of 30 PubChem actives in DS4 contains an average of only 1.16 molecules per scaffold, and the data set hence provides an obvious basis for probing further the effectiveness of different weighting functions for searching structurally diverse sets of actives. The search results for DS4 are shown in Table 8 and are rather different from those for DS3. Results in Table 8 show that SQR1 is the best performing weighting function across the 17 activity classes in terms of mean recall, with OKA and OKA1 also performing well. The overall ranking of the 10 functions based on the Kendall coefficient in Table 9 for the top 1% for DS4 is SQR1 > OKA > OKA1 > AUG1 > SQR > SMO > STD1 > SMO1 > STD = AUG.

Table 9 gives the level of statistical significance and the associated ranking for the set of results in Tables 5−8. It will be seen that there is a high degree of commonality in the rankings, with modified weighting functions (e.g., OKA1, SMO1, SQR1, and AUG1), in particular providing a level of performance that is generally superior to the other weighting functions tested here (unmodified weighting functions). The only obvious exception is the DS2 and DS4 top 5% searches, where the new weighting functions are ranked second. In addition, the results in Table 10 show that the higher values (i.e., corresponding to greater number of high-effectiveness searches) tend to occur in the right-hand side of this table (corresponding to weighting functions STD1, OKA1, AUG1, SQR1, and SMO1).

Now and based on the results presented here, we can draw the following conclusions: First, introducing new weighting functions (using new weighting schemes) can provide significantly more effective screening. Second, no single best weighting function can offer a consistently high level of screening performance in all circumstances (different databases with different level of diversities). This finding is in line with previous studies by Sheridan[3] and Abdo et al.[13] In their works they found that no single best method can always yield a high level of screening performance in all circumstances. Third, the OKA1 function can be considered (on average) as the best function from others.

CONCLUSION

In more recent studies, Bayesian networks for ligand-based virtual screening have been investigated. The main aim for the network model is to improve the retrieval effectiveness in searching chemical databases. The success or failure of the network model depends on the fragment weighting functions which combine one or more fragment weighting schemes. In this paper we have developed a new fragment weighting scheme and further investigate five different weighting functions which have been originally used in the information retrieval area. Simulated virtual screening experiments with the MDDR and MUV data sets show that the use of weighting functions which combine the new weighting scheme can provide significantly more effective screening when compared with the use of current weighting schemes. We conclude that developing a new fragment weighting scheme can provide a very simple way of increasing the screening effectiveness of BIN.

ACKNOWLEDGMENT

We thank Prof. Peter Willett and referees for valuable comments on an initial draft of this manuscript.

REFERENCES AND NOTES

(1) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today* **2007**, *12*, 225–233.

(2) Geppert, H.; Vogt, M.; Bajorath, J. 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) Sheridan, R. P. Chemical similarity searches: When is complexity justified. *Expert Opin. Drug Discovery* **2007**, *2*, 423–430.

(4) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.

(5) Willett, P. Similarity Methods in Chemoinformatics. *Ann. Rev. Inf. Sci. Technol.* **2009**, *43*, 3–71.

(6) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*, 2nd ed.; Kluwer: Dordrecht, The Netherlands, 2007.

(7) Turtle, H.; Croft, W. B. In *Inference Networks for Document Retrieval*, Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Brussels, Belgium, September 5−7, 1990 Association for Computing Machinery: New York, 1990.

(8) Turtle, H.; Croft, W. B. Evaluation of an Inference Network-Based Retrieval Model. *ACM Trans. Inf. Syst.* **1991**, *9*, 187–222.

(9) Abdo, A.; Salim, N. In *Inference Networks for Chemical Similarity Searching*, Proceedings of the International Conference on Advanced Computer Theory and Engineering, Phuket, Thailand, December 20−22, 2008; IEEE Computer Society: Los Alamitos, CA, 2008; pp 408−412.

(10) Abdo, A.; Salim, N. Similarity-Based Virtual Screening with a Bayesian Inference Network. *ChemMedChem* **2009**, *4*, 210–218.

(11) 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.

(12) Chen, B.; Mueller, C.; Willett, P. Evaluation of a Bayesian inference network for ligand-based virtual screening. *J. Cheminf.* **2009**, *1* (5); http://www.jcheminf.com/content/1/1/5; DOI: 10.1186/1758-2946-1-5. Accessed November 19, 2010.

(13) 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.

(14) Abdo, A. Similarity-Based Virtual Screening Using Bayesian Inference Network for Searching Chemical Database. Ph.D. Thesis, Universiti Teknologi Malaysia, Malaysia, 2009.

(15) Willett, P. Textual and chemical information retrieval: different domains but similar algorithms. *Inf. Research* **2000**, *5*; http://informationr.net/ir/5-2/paper69.html. Accessed November 19, 2010.

(16) Abdo, A.; Salim, N. In *Molecular Similarity Searching Using Inference Network*, Proceedings of the 237th ACS National Meeting of the American Chemical Society, Salt Lake City, UT, March 22−26, 2009; American Chemical Society: Washington, DC, 2009.

(17) Callan, J. P.; Croft, W. B.; Broglio, J. TREC and TIPSTER experiments with InQuery. *Inf. Process. Manage.* **1995**, *31*, 327–343.

(18) Robertson, S. E.; Walker, S. In *Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval*, Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, July 3−6, 1994; Association for Computing Machinery: New York, 1994; pp 232−241.

(19) Robertson, S. E.; Walker, S.; Hancock-Beaulieu, M. M. Large test collection experiments on an operational, interactive system: Okapi at TREC. *Inf. Process. Manage.* **1995**, *31*, 345–360.

(20) Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.* **1988**, *24*, 513–523.

(21) Chisholm, E.; Kolda, T. G. *New Term Weighting Formulas for the Vector Space Method in Information Retrieval*; Technical Report; Oak Ridge National Laboratory: Oak Ridge, TN, 1999.

(22) Metzler, D.; Croft, W. B. Combining the language model and inference network approaches to retrieval. *Inf. Process. Manage.* **2004**, *40*, 735–750.

(23) MDL Drug Data Report; Symyx Technologies: San Diego, CA; http://www.symyx.com/products/databases/bioactivity/mddr/index.jsp. Accessed November 19, 2010.

(24) *Pipeline Pilot*; Accelrys Software Inc.: San Diego, CA, 2008.

(25) 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.

(26) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, *49*, 169–184.

(27) Siegel, S.; Castellan, N. J. *Nonparametric Statistics for the Behavioral Sciences*; McGraw-Hill: New York, 1988.

CI100232H