Ligand-Based Virtual Screening Using Bayesian Networks

Ammar Abdo,*,† Beining Chen,‡,§ Christoph Mueller,‡,|| Naomie Salim,† and Peter Willett‡,||

Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310, Skudai, Malaysia, Krebs Institute for Biomolecular Research, and Departments of Chemistry, Information Studies, University of Sheffield, Sheffield S10 2TN, U.K.

Received March 5, 2010

A Bayesian inference network (BIN) provides an interesting alternative to existing tools for similarity-based virtual screening. The BIN is particularly effective when the active molecules being sought have a high degree of structural homogeneity but has been found to perform less well with structurally heterogeneous sets of actives. In this paper, we introduce an alternative network model, called a Bayesian belief network (BBN), that seeks to overcome this limitation of the BIN approach. Simulated virtual screening experiments with the MDDR, WOMBAT and MUV data sets show that the BIN and BBN methods allow effective screening searches to be carried out. However, the results obtained are not obviously superior to those obtained using a much simpler approach that is based on the use of the Tanimoto coefficient and of the square roots of fragment occurrence frequencies.

INTRODUCTION

Virtual screening is the name given to a range of computational tools for searching chemical databases to identify molecules that have a high probability of exhibiting activity against a specific biological target. These tools are widely used to reduce the costs of drug discovery by filtering out that large fraction of a typical database that is unlikely to yield positive results in conventional biological screening. There are two main types of virtual screening: structure-based approaches (e.g., de novo design and ligand-protein docking) can be used when the 3D structure of the biological target is available, and ligand-based approaches (e.g., similarity searching, pharmacophore mapping, and machine learning) are applicable in the absence of such structural information. The choice of virtual screening method depends on the amount and type of data that are available,[1–3] with the methods differing considerably in the associated computational requirements; for example, 3D virtual screening methods may require the generation of low-energy conformers for all of the molecules in the database that is being searched.[4]

Similarity searching is the simplest and one of the most widely used tools for ligand-based virtual screening, since it requires just a single known bioactive molecule, or reference structure, as the starting-point for a database search. The basic idea underling similarity searching is the similar property principle, which states that structurally similar molecules will exhibit similar physicochemical and biological properties.[5] Over the years, many ways of measuring the structural similarity of molecules have been introduced.[4,6–9] The most common approach, which we study in this paper, 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, normally the Tanimoto coefficient.[1,6]

In this paper, we discuss alternative approaches to the calculation of fingerprint-based molecular similarities that have been inspired by research in textual information retrieval on the use of inference networks for database searching.[10–12] There are several analogies between textual information retrieval and chemoinformatics,[13] and these led to recent work by Abdo and Salim[14–16] that developed a ligand-based virtual screening method that uses a Bayesian inference network (BIN) and 2D fingerprints. Experiments with a subset of the MDL Drug Data Report (MDDR)[17] database demonstrated that the BIN provided an interesting alternative to existing tools for ligand-based virtual screening. This was especially so when the active molecules being sought had a high degree of structural homogeneity, when the BIN substantially outperformed a conventional, Tanimoto-based similarity searching system. Similar results were obtained by Chen et al.,[18] who used a BIN to search the MDDR and World Of Molecular Bioactivity (WOMBAT)[19] databases. However, both of these studies found that the effectiveness of the BIN was much less when structurally heterogeneous sets of actives were being sought.

This paper reports the use of a new form of Bayesian network for similarity-based virtual screening that has been developed with the aim of overcoming the limitations of the BIN approach. This new approach is based on work by Ribeiro and Muntz that interprets the probabilities as degrees of belief that can be used to rank searches in information retrieval.[11] Here, we use a Bayesian network to provide a formal framework for these degrees of belief, and we hence refer to our new approach as a Bayesian belief network (BBN). In this paper, we describe the BBN and compare it

* To whom correspondence should be addressed. E-mail: Ammar_utm@yahoo.com. Phone: 006-017-7425041. Fax: 006-07-5533210.
† Universiti Teknologi Malaysia.
‡ Krebs Institute for Biomolecular Research.
§ Department of Chemistry, University of Sheffield.
|| Department of Information Studies, University of Sheffield.

10.1021/ci100090p
Published on Web 05/26/2010

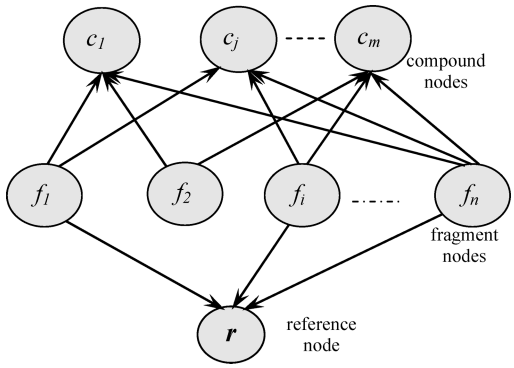

Figure 1. Bayesian belief network (BBN) model.

with two other approaches to virtual screening using three standard chemical data sets.

MATERIALS AND METHODS

This study has compared the retrieval results obtained using three different similarity-based screening systems. The first system was based on BBN. The second screening system was based on the BIN described by Chen et al.,[18] specifically using the Okapi (OKA) weight that performed best in their experiments. The third screening system was based on the Tanimoto (TAN) coefficient, which has been used for ligand-based virtual screening for many years and which can hence be considered as a reference standard. In what follows, we give a detailed description of the BBN approach, and briefer accounts of the previously described BIN and TAN approaches.

Bayesian Belief Network Model. We have noted previously the close relationship that exists between many of the methods that are used for textual information retrieval and for chemoinformatics[13] and it is this relationship that motivated our interest in the application of BIN and BBN to virtual screening. Specifically, we suggest here that the BBN model introduced by Ribeiro and Muntz[11] can be applied to ligand-based virtual screening by replacing the index terms, the document database and the textual query in their model by substructural fragments, a database of 2D structures and a reference structure, respectively.

The BBN model is shown in Figure 1 and consists of three types of nodes: fragment nodes $\boldsymbol{f}$ as roots, and compound nodes $\boldsymbol{c}$ and a reference-structure node $\boldsymbol{r}$ as leaves (where the roots of the network are the nodes without parent nodes, and the leaves are the nodes without child nodes). Each compound node has one or more fragment nodes as parents, as does the reference-structure node. Both the database structures and the reference structure are described using a set, S, of 2D fragments, the presence/absence of which is encoded in a fingerprint. For simplicity of processing, the fragment occurrences are assumed to be statistically independent of each other, an assumption that was also adopted in our previous BIN work. Each network node is binary-valued, taking one of two values from S.

The set S can be modeled by n independent network nodes, with each node associated with a binary random variable that is set to "1" if that fragment node has been assigned to a specific compound node or reference-structure node. However, the calculation is carried out only for those nodes representing fragments common to the reference-structure node and a compound node. Each fragment node has a prior probability associated with it that describes the probability of observing that fragment: this probability will be set to $1/n$, where n is the total number of fragments.

The reference-structure is modeled by a network node that has one or more fragment nodes as parents, and that contains a specification of the conditional probability associated with the node given its set of parent fragment nodes. This specification incorporates the effect of any weighting scheme associated with the reference node; weighting schemes are discussed further below. An entirely comparable representation is used for each compound in the database that is to be searched. The total number of fragment nodes corresponds to the length of the fingerprint used to characterize the compounds and reference structures, and we represent the reference-structure and compound nodes as fragment vectors. Let $\vec{c} = (f_1, ..., f_n)$ be the vector representing the compound $\boldsymbol{c}$ and $\vec{r} = (f'_1, ..., f'_n)$ be the vector representing the reference structure $\boldsymbol{r}$.

To complete our belief network we need to estimate the strength of the relationships represented by the network, and this estimation process involves estimating and encoding a set of conditional probability distributions. Specifically, for any non-root node A in the network, where A has a set of parent nodes $\{P_1, P_2, ..., P_n\}$, we must estimate the probability $P(A|P_1,P_2, ..., P_n)$, that is, we need to specify the conditional probability for the non-root nodes $P(c|\vec{f})$ and $P(r|\vec{f'})$. As shown in our previous studies,[15,18] these probabilities can be estimated by many different type of weighting schemes. These weighting schemes has been used previously in information retrieval to model information about the occurrences of textual keywords, but has been modified here to model information about the occurrences of substructural fragments. Here, we use the following weighting scheme to estimate the probabilities for each fragment in a compound or reference-structure vector. This weighting scheme was originally used in InQuery system[20,21] and has been used successfully in our previous studies.[15,18]

$$P(c_j|\vec{f}) = \alpha + (1 - \alpha) \times \frac{ff_{ij}}{\max ff_j} \times \frac{\log\left(\frac{m + 0.5}{cf_i}\right)}{\log(m + 1.0)} \quad (1)$$

$$P(r|\vec{f'}) = \alpha + (1 - \alpha) \times \frac{ff_{ir}}{\max ff_r} \times \frac{\log\left(\frac{m + 0.5}{cf_i}\right)}{\log(m + 1.0)} \quad (2)$$

where α is a constant (which was set to 0.4 in our experiments), ff_{ij} and ff_{ir} are the frequency of the ith fragment within the jth compound and the rth reference structure, respectively, $\max ff_j$ and $\max ff_r$ are the maximum frequency of fragment occurrence in the jth compound and the rth reference structure, respectively, cf_i is the number of compounds containing the ith fragment, and m is the number of compounds in the collection. 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

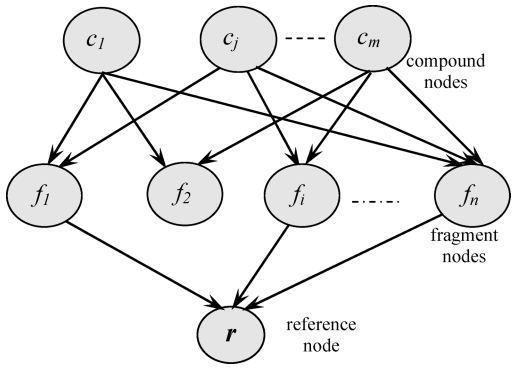

Figure 2. Bayesian inference network (BIN) model.

the weighted-max (WMAX) or weighted-sum (WSUM) operators as described previously for the BIN model.[22]

To produce a ranking of the compounds in the collection with respect to a given reference structure, the Ribeiro−Muntz model would suggest the use of the cosine coefficient, which measures the cosine of the angle between the weighted reference-structure and compound vectors. This cosine similarity is

$$bel(\vec{c},\vec{r}) = \frac{\vec{c}\bullet\vec{r}}{|\vec{c}|\times|\vec{r}|} = \frac{\sum_{i=1}^{n} c_i \times r_i}{\sqrt{\sum_{i=1}^{n} c_i^2} \times \sqrt{\sum_{i=1}^{n} r_i^2}} \quad (3)$$

However, we found that this approach, which is used very extensively in textual information retrieval, performed poorly here, and that superior results were obtained using

$$bel(\vec{c},\vec{r}) = \frac{c_j}{cl_j} \times \sum_{i=1}^{n} (1 - |p_{ij} - p_i|) \quad (4)$$

Here, c_j is the set of fragments in common between the jth compound and the reference-structure ($c_j \leq n$), cl_j is the number of the unique fragments assigned to the jth compound, p_{ij} and p_i are the estimated probabilities at the ith fragment node for the jth compound and for the reference-structure, respectively (as computed using eqs 1 and 2).

The poor performance of the cosine coefficient here has several possible explanations. First, the calculation has been done here only for the common fragment nodes common to the reference-structure node and each compound node, rather than for all of the nodes. Second, the values of reference and compound vectors are represented only the common fragments to the reference and compound which calculated in similar way. In addition, these two points represent the rationale to assess the similarity between reference and compound vectors using eq 4 rather than eq 3.

Bayesian Inference Network Model. The second similarity search system (BIN) was used as described in detail by Chen et al.[18] and as shown in Figure 2. The difference between the two approaches (BIN and BBN) arises from differences in the nature of the network topology and the calculation of the belief values, as discussed below. In addition, the two approaches use a different ranking strategy: that described by Chen et al. for the BIN and that defined by eq 4 in BBN.

We consider first the nature of the networks. Comparison of Figures 1and 2 suggests a clear resemblance, but there are important differences. In BIN, the root and leaf nodes represent the database compounds and the reference-structure, respectively; in BBN, the root nodes represent substructural fragments and the leaf nodes represent the reference-structure and the database compounds. Consequently, the information in BIN propagates from the compound node (root) toward the reference-structure node (leaf), whereas the information in BBN propagates from the fragment nodes (root) toward the reference-structure node and compound nodes (leaves). In addition, the ranking strategies in the two models are different.

Turning now to the calculation of the belief values, BIN differs from BBN in three ways. First, Chen et al. found that the Okapi BM25 belief function was found to perform best of those tested, and this has been used in the BIN implementation tested here. Second, a simple #SUM-operator is used to aggregate the probability scores for each database compound. Third, the aggregation includes all of the n fragments, and not just the ones for which $ff > 0$ as in the BBN approach.

Tanimoto-Based Similarity Searching. The third similarity search system (TAN) used the continuous form of the Tanimoto coefficient, which is applicable to non-binary data such as used here. The similarity score $S_{A,B}$ for molecules A and B was calculated using eq 5.

$$S_{A,B} = \frac{\sum_{i=1}^{n} f_i f'_i}{\sum_{i=1}^{n} (f_i)^2 + \sum_{i=1}^{n} (f'_i)^2 - \sum_{i=1}^{n} f_i f'_i} \quad (5)$$

Equation 5 has been widely used for chemical similarity searching. However, a detailed study of fragment weighting schemes has recently suggested that superior screening performance is obtained if the square roots of the fragment occurrence frequencies are used in the Tanimoto coefficient, rather than the unmodified frequencies,[23] and we have hence carried out experiments in which the raw fragment frequencies in the BIN, BBN, and TAN similarity measures are replaced by the square roots of those frequencies: these searches will be referred to as BINS, BBNS, and TANS, respectively.

Simulated Virtual Screening Experiments. Our first sets of virtual screening experiments used the popular MDDR and WOMBAT databases that have been employed in our previous studies of Bayesian networks,[15,18] with the two databases containing 102 516 and 138 127 molecules, respectively. All molecules in both databases were 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 and one data set (DS4) was chosen (as described by Gardiner et al.[26]) from the WOMBAT 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

Table 1. MDDR Activity Classes for DS1 Data Set

activity index	activity class	active molecules	pairwise similarity mean	pairwise similarity SD
31420	renin inhibitors	1130	0.573	0.106
71523	HIV protease inhibitors	750	0.446	0.122
37110	thrombin inhibitors	803	0.419	0.127
31432	angiotensin II AT1 antagonists	943	0.403	0.101
42731	substance P antagonists	1246	0.339	0.106
06233	5HT3 antagonists	752	0.351	0.116
06245	5HT reuptake inhibitors	359	0.345	0.122
07701	D2 antagonists	395	0.345	0.103
06235	5HT1A agonists	827	0.343	0.104
78374	protein kinase C inhibitors	453	0.323	0.142
78331	cyclooxygenase inhibitors	636	0.268	0.093

Table 2. MDDR Activity Classes for DS2 Data Set

activity index	activity class	active molecules	pairwise similarity mean	pairwise similarity SD
07707	adenosine (A1) agonists	88	0.542	0.124
07708	adenosine (A2) agonists	71	0.536	0.137
31420	renin inhibitors	1130	0.459	0.119
42710	CCK agonists	79	0.452	0.099
64100	monocyclic β-lactams	76	0.549	0.084
64200	cephalosporins	1312	0.501	0.098
64220	carbacephems	73	0.487	0.099
64500	carbapenems	896	0.457	0.124
64350	tribactams	74	0.548	0.150
75755	vitamin D analogous	279	0.574	0.105

Table 3. MDDR Activity Classes for DS3 Data Set

activity index	activity class	active molecules	pairwise similarity mean	pairwise similarity SD
09249	muscarinic (M1) agonists	848	0.206	0.098
12455	NMDA receptor antagonists	1311	0.199	0.090
12464	nitric oxide synthase inhibitors	377	0.189	0.086
31281	dopamine β-hydroxylase inhibitors	95	0.229	0.076
43210	aldose reductase inhibitors	882	0.232	0.096
71522	reverse transcriptase inhibitors	519	0.218	0.095
75721	aromatase inhibitors	513	0.229	0.117
78331	cyclooxygenase inhibitors	636	0.220	0.107
78348	phospholipase A2 inhibitors	704	0.224	0.111
78351	lipoxygenase inhibitors	2555	0.224	0.110

(i.e., structurally diverse). The DS2 data set contains 10 homogeneous activity classes and the DS3 data set 10 heterogeneous activity classes. The 14 activity classes in the WOMBAT DS4 data set are analogous to DS1 in that they include both homogeneous and heterogeneous activity classes. The four data sets are listed 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 Unity 2D fingerprints (available from Tripos Inc. at http://www.tripos.com).

Further experiments involved the maximum unbiased validation (MUV) data set (Table 5) reported recently by Rohrer and Baumann.[27] This contains 17 activity classes,

Table 4. WOMBAT Activity Classes for DS4 Data Set

activity index	activity class	active molecules	pairwise similarity mean	pairwise similarity SD
31420	renin inhibitors	474	0.592	0.109
78374	protein kinase C inhibitors	142	0.565	0.277
78432	matrix metalloprotease inhibitors	694	0.444	0.148
31432	angiotensin II AT1 antagonists	724	0.443	0.131
71523	HIV protease inhibitors	1128	0.442	0.146
42731	substance P antagonists	558	0.427	0.127
37110	thrombin inhibitors	421	0.418	0.144
06235	5HT1A antagonists	592	0.399	0.134
37121	factor Xa inhibitors	842	0.394	0.124
06233	5HT3 antagonists	220	0.377	0.175
09221	acetylcholine esterase inhibitors	503	0.373	0.155
07701	D2 antagonists	910	0.367	0.116
28310	phosphodiesterase inhibitors	596	0.359	0.136
78331	cyclooxygenase inhibitors	965	0.324	0.139

Table 5. MUV Activity Classes for DS5 Data set

activity index	activity class	active molecules	pairwise similarity mean	pairwise similarity SD
466	S1P1 rec. (agonists)	30	0.285	0.072
548	PKA (inhibitors)	30	0.293	0.094
600	SF1 (inhibitors)	30	0.288	0.085
644	rho-kinase2 (inhibitors)	30	0.267	0.096
652	HIV RT-RNase (inhibitors)	30	0.258	0.079
689	Eph rec. A4 (inhibitors)	30	0.267	0.074
692	SF1 (agonists)	30	0.247	0.062
712	HSP 90 (inhibitors)	30	0.260	0.073
713	ER-a-Coact. bind. (inhibitors)	30	0.261	0.073
733	ER-β-Coact. bind. (inhibitors)	30	0.266	0.072
737	ER-a-Coact. bind. (potentiators)	30	0.297	0.068
810	FAK (inhibitors)	30	0.277	0.082
832	cathepsin G (inhibitors)	30	0.319	0.131
846	FXIa (inhibitors)	30	0.281	0.109
852	FXIIa (inhibitors)	30	0.295	0.099
858	D1 rec. (allosteric modulators)	30	0.247	0.092
859	M1 rec. (allosteric inhibitors)	30	0.275	0.072

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 the data set hence provides a much stiffer test of screening effectiveness than the other data sets studied here. An inactive molecule in a search of one of the MDDR data sets (DS1−DS3) is one that has not been allocated the appropriate database activity descriptor. The inactive molecules in the WOMBAT and MUV data sets (DS4 and DS5) are as described by Gardiner et al.[26] and by Rohrer and Baumann.[27] The molecules here were again represented by ECFC4 fingerprints.

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 molecules, 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

The results for the searches of DS1−DS5 are shown in Tables 6−10, respectively, using cutoffs of both 1% and 5%.

Table 6. Retrieval Results for Data Set DS1

	raw frequencies						square root of frequencies					
	TAN		BIN		BBN		TANS		BINS		BBNS	
activity index	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%
31420	55.84	85.49	62.44	94.25	57.78	91.21	63.19	93.69	66.60	94.06	57.78	91.21
71523	22.26	42.76	24.11	47.05	24.90	46.90	29.00	53.95	25.58	45.58	24.86	46.88
37110	12.54	24.11	21.51	46.15	21.74	43.95	18.82	38.79	24.23	47.24	21.74	43.93
31432	33.37	68.20	41.56	80.15	39.55	77.08	41.91	83.58	41.24	79.46	39.54	77.08
42731	16.25	32.81	19.11	30.83	14.55	27.74	19.69	35.43	18.01	29.59	14.55	27.75
06233	14.22	27.01	17.19	27.56	19.04	32.81	17.55	28.91	17.51	28.92	19.04	32.83
06245	10.61	22.90	11.41	23.02	11.70	23.89	11.98	24.83	12.09	24.12	11.69	23.87
07701	08.91	23.10	09.68	21.82	09.44	21.49	9.22	22.99	9.84	22.27	9.46	21.49
06235	11.87	24.54	13.72	25.11	13.80	25.94	13.42	26.77	13.71	25.19	13.80	25.95
78374	16.75	24.26	13.83	23.08	14.39	22.10	16.96	24.39	13.22	22.38	14.40	22.10
78331	8.05	16.83	7.69	14.14	8.35	16.24	8.54	16.32	7.60	14.64	8.36	16.22
av	19.15	35.64	22.02	39.38	21.39	39.03	22.75	40.88	22.69	39.40	21.38	39.03
SD	14.12	21.77	16.30	25.83	14.94	24.37	16.48	25.71	17.34	25.55	14.94	24.37

Table 7. Retrieval Results for Data Set DS2

	raw frequencies						square root of frequencies					
	TAN		BIN		BBN		TANS		BINS		BBNS	
activity index	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%
07707	78.30	91.08	98.64	100.00	97.73	100.00	99.03	99.94	99.55	100.00	97.73	100.00
07708	74.01	88.52	96.97	100.00	96.13	100.00	98.87	100.00	97.89	100.00	96.13	100.00
31420	46.44	77.60	46.76	92.19	43.90	82.81	48.50	88.07	53.71	91.84	43.90	82.81
42710	57.22	67.59	53.99	68.73	55.06	67.78	59.37	69.62	58.48	68.67	55.06	67.78
64100	93.22	97.89	95.92	98.55	97.30	98.49	96.38	99.41	96.05	98.49	97.30	98.49
64200	63.39	89.82	75.33	99.55	73.68	99.55	72.67	98.89	75.35	99.57	73.68	99.55
64220	73.56	92.05	75.62	99.66	83.22	99.66	79.11	99.66	76.16	99.73	83.22	99.66
64500	60.75	74.98	88.57	99.01	89.18	99.16	84.69	97.62	89.29	99.11	89.18	99.16
64350	76.69	90.34	100.00	100.00	100.00	100.00	99.59	100.00	100.00	100.00	100.00	100.00
75755	95.99	98.78	99.64	99.64	99.64	99.64	99.62	99.64	99.64	99.64	99.64	99.64
av	71.96	86.87	83.14	95.73	83.58	94.71	83.78	95.28	84.61	95.70	83.58	94.71
SD	15.51	10.17	19.64	09.78	19.97	10.84	18.58	09.72	17.67	09.82	19.97	10.84

Table 8. Retrieval Results for Data Set DS3

	raw frequencies						square root of frequencies					
	TAN		BIN		BBN		TANS		BINS		BBNS	
activity index	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%
09249	25.09	40.21	17.89	28.94	21.57	34.10	22.89	35.93	19.68	30.50	21.63	34.12
12455	7.70	19.08	6.28	12.17	7.47	15.47	7.72	17.41	6.72	13.05	7.47	15.47
12464	9.02	14.56	8.73	16.25	11.60	18.10	9.83	15.31	10.03	18.57	11.61	18.10
31281	27.53	44.00	26.26	34.95	31.37	43.26	27.05	40.74	27.53	35.95	31.47	43.32
43210	11.10	26.37	10.61	19.31	12.89	23.70	12.13	24.89	11.34	20.36	12.89	23.70
71522	2.35	6.28	3.29	7.04	3.54	7.57	3.22	7.24	3.40	7.50	3.53	7.57
75721	24.02	28.97	22.75	28.52	23.73	30.68	24.34	28.80	23.31	29.60	23.72	30.65
78331	6.27	15.79	5.10	10.08	5.88	12.68	6.25	13.22	5.25	10.94	5.87	12.69
78348	4.69	13.16	3.60	11.25	4.84	14.35	5.17	15.82	3.59	11.37	4.85	14.36
78351	4.31	10.55	4.12	9.58	4.84	12.59	4.30	10.52	4.38	10.23	4.84	12.60
av	12.21	21.90	10.86	17.81	12.77	21.25	12.29	20.99	11.52	18.81	12.79	21.26
SD	9.57	12.66	8.44	9.75	9.61	11.41	9.04	11.16	8.86	10.01	9.64	11.42

The left-hand part of each table contains the results for the three basic searches (TAN, BIN, and BBN), based on the use of the raw fragment frequencies of occurrence; the right-hand part of each table contains the corresponding results when the square roots of the frequencies are used. Each row in a table corresponds to one activity class, and the two bottom rows in a table correspond to the mean (AVG) and the standard deviation (SD) when averaged over all of the activity classes for a data set.

Visual inspection of the recall values in Tables 6−10 enables one to make comparisons between the effectiveness of the various search methods. However, a more quantitative approach is possible using the Kendall *W* test of concordance.[28] This was developed to quantify the level

Table 9. Retrieval Results for Data Set DS4

	raw frequencies						square root of frequencies					
	TAN		BIN		BBN		TANS		BINS		BBNS	
activity index	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%
06235	20.62	44.13	23.31	47.08	19.92	42.59	21.88	46.89	23.51	46.00	19.91	42.62
06233	21.82	36.75	24.39	34.36	25.75	37.04	25.41	36.73	24.55	34.43	25.80	37.02
09221	10.67	18.97	17.59	27.23	16.32	25.88	14.12	23.62	18.09	27.25	16.32	25.89
31432	30.87	54.38	52.13	73.57	45.12	68.14	48.97	73.30	52.13	73.96	45.14	68.13
78331	28.85	41.80	29.11	38.48	28.92	40.28	30.68	42.95	29.39	38.57	28.93	40.29
07701	13.69	33.87	14.23	33.14	13.20	31.95	14.25	36.27	14.50	33.29	13.21	31.97
37121	18.34	33.69	22.43	35.80	20.29	31.40	21.98	36.03	22.57	36.21	20.23	31.38
71523	20.78	41.71	25.24	48.61	25.35	47.66	25.87	51.90	28.03	49.89	25.34	47.67
78432	20.85	35.74	28.38	47.64	30.44	49.67	30.39	48.59	28.34	48.46	30.46	49.66
28310	13.31	21.59	20.97	29.23	20.57	29.48	19.06	26.81	21.54	30.09	20.60	29.47
78374	59.09	62.71	58.38	63.24	57.43	63.66	59.08	64.12	57.71	62.99	57.43	63.66
31420	59.38	81.56	70.41	93.56	59.53	89.56	70.20	91.99	76.41	95.56	59.52	89.57
42731	24.84	37.02	29.68	38.50	25.41	40.26	29.15	39.95	29.65	38.72	25.43	40.27
37110	18.90	34.54	22.17	41.60	23.04	41.33	23.69	44.42	25.70	44.96	23.03	41.31
av	25.86	41.32	31.32	46.58	29.38	45.64	31.05	47.40	32.29	47.17	29.38	45.64
SD	15.19	16.11	16.66	18.60	14.44	17.50	16.72	18.40	17.42	18.88	14.44	17.50

Table 10. Retrieval Results for Data Set DS5

	raw frequencies						square root of frequencies					
	TAN		BIN		BBN		TANS		BINS		BBNS	
activity index	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%	top 1%	top 5%
AC466	5.55	10.22	6.33	10.44	5.89	9.33	6.00	8.45	6.22	10.33	5.89	9.33
AC548	11.67	23.33	14.89	27.22	13.44	22.78	13.22	25.00	14.78	26.89	13.44	22.78
AC600	6.44	14.22	6.33	12.89	6.22	10.89	6.22	12.11	6.55	12.78	6.22	10.89
AC644	11.00	21.22	11.00	19.67	9.44	20.00	11.00	20.78	10.89	19.55	9.44	20.00
AC652	6.33	13.78	7.00	11.67	5.89	10.11	6.44	11.00	6.56	12.00	5.89	10.11
AC689	5.00	11.56	7.33	13.22	7.11	11.89	6.44	13.00	7.56	12.67	7.00	11.89
AC692	4.33	9.00	5.33	9.22	4.78	9.22	4.89	8.78	5.22	9.11	4.78	9.22
AC712	8.22	16.44	8.22	16.45	6.89	12.44	8.00	14.67	7.89	16.45	6.89	12.55
AC713	6.00	10.44	5.89	9.00	6.22	9.89	6.22	9.44	6.00	9.11	6.22	9.89
AC733	5.22	9.44	6.67	10.11	6.55	9.44	6.33	9.56	6.55	9.78	6.55	9.44
AC737	5.78	15.78	5.11	12.00	5.22	10.22	5.33	11.67	5.33	11.56	5.22	10.22
AC810	6.00	11.78	6.78	13.33	6.22	11.11	6.11	11.00	6.78	13.00	6.11	11.00
AC832	9.67	17.00	12.55	20.44	13.11	22.33	12.33	20.33	12.67	21.11	13.11	22.33
AC846	10.56	22.45	13.11	26.11	13.22	28.11	12.00	27.00	13.78	27.22	13.11	28.11
AC852	11.67	19.67	13.78	23.11	13.78	22.67	14.11	22.33	14.11	23.67	13.67	22.67
AC858	5.67	11.11	5.11	9.11	5.22	8.56	6.00	9.89	5.11	9.22	5.22	8.56
AC859	5.33	12.67	4.89	9.44	4.55	9.22	4.66	11.44	4.67	9.00	4.55	9.22
av	7.32	14.71	8.25	14.91	7.87	14.01	7.96	14.50	8.27	14.91	7.84	14.01
SD	2.55	4.65	3.39	6.13	3.34	6.35	3.18	6.08	3.49	6.37	3.32	6.35

of agreement between multiple sets of rankings of the same set of objects: here, and in previous work,[18] we have used this approach to rank the effectiveness of different search methods. The basic form of the W statistic is

$$\frac{12\sum_{i=1}^{N}(R_i - R)}{N^3 - N} \tag{6}$$

where R_i is the mean of the ranks assigned to the ith object and R is the grand mean of the ranks assigned to all N objects. In the present context, we consider each activity class in a data set as a judge that is ranking the different searches in order of decreasing screening effectiveness (as measured by the recall value). Thus, the raw recall data (i.e., that summarized in Tables 6−10) is converted to ranks so that, for example, the most effective search gets a rank of one. The ranks are then used to compute the Kendall statistic, the statistical significance of which can be tested using the χ^2 distribution since for $N > 7$

$$\chi^2 = k(N - 1)W \tag{7}$$

with $N - 1$ degrees of freedom (alternatively, Siegel and Castellan provide a table of critical values for W when $N \leq 7$[28]). If a statistically significant value is obtained in this test (for which we used the 0.01 or 0.05 levels of statistical significance), then Siegel and Castellan suggest that one can obtain an overall ranking of the set of objects that is being judged.

The results of the Kendall analyses are reported in Tables 11−15; in each case, the left- and right-hand portions of the

Table 11. Kendall *W* Test Results (Top 1% and Top 5%) Using DS1

top 1%				top 5%			
W	*p*	method	rank	*W*	*p*	method	rank
0.271	<0.01	TANS	2.45	0.199	>0.05	TANS	2.00
		BINS	2.91			BINS	3.27
		BBNS	3.23			BIN	3.45
		BBN	3.41			BBN	3.95
		BIN	3.73			BBNS	3.95
		TAN	5.27			TAN	4.36

Table 12. Kendall *W* Test Results (Top 1% and Top 5%) Using DS2

top 1%				top 5%			
W	*p*	method	rank	*W*	*p*	method	rank
0.378	<0.01	BINS	2.10	0.584	<0.01	BINS	2.65
		BBN	3.15			BIN	2.80
		TANS	3.20			BBN	3.05
		BBNS	3.45			BBNS	3.25
		BIN	3.60			TANS	3.25
		TAN	5.50			TAN	6.00

Table 13. Kendall *W* Test Results (Top 1% and Top 5%) Using DS3

top 1%				top 5%			
W	*p*	method	rank	*W*	*p*	method	rank
0.468	<0.01	BBNS	2.35	0.445	<0.01	BBNS	2.35
		BBN	2.45			TAN	2.80
		TANS	2.90			BBN	2.85
		TAN	3.40			TANS	3.10
		BINS	4.20			BINS	4.20
		BIN	5.70			BIN	5.70

Table 14. Kendall *W* Test Results (Top 1% and Top 5%) Using DS4

top 1%				top 5%			
W	*p*	method	rank	*W*	*p*	method	rank
0.416	<0.01	BINS	1.93	0.211	<0.01	TANS	2.57
		BIN	2.86			BINS	2.64
		TANS	2.86			BIN	3.57
		BBNS	3.93			BBNS	3.61
		BBN	4.07			BBN	3.68
		TAN	5.36			TAN	4.93

Table 15. Kendall *W* Test Results (Top 1% and Top 5%) Using DS5

top 1%				top 5%			
W	*p*	method	rank	*W*	*p*	method	rank
0.163	<0.05	BINS	2.56	0.120	>0.05	BIN	2.62
		BIN	2.68			BINS	3.09
		TANS	3.50			TAN	3.18
		BBN	3.85			TANS	3.79
		BBNS	4.18			BBN	4.15
		TAN	4.24			BBNS	4.18

tables describe the top 1% and top 5% rankings, respectively. In each table section, the columns show the value of the coefficient, the associated probability, the similarity method, and the mean rank of that method when the methods are ranked in decreasing order of screening effectiveness (if two methods have the same mean rank then they are ordered on the basis of the mean recall, that is, the AVG values from the main tables of results).

Table 16. Mean Numbers of Top 1% Compounds Common to a Pair of Search Rankings

	TAN	BIN	BBN	TANS	BINS	BBNS
TAN	1025.00	573.16	574.69	692.98	580.09	574.74
BIN		1025.00	841.66	816.85	934.51	841.51
BBN			1025.00	825.58	873.85	1024.18
TANS				1025.00	814.79	825.37
BINS					1025.00	873.73
BBNS						1025.00

We shall use the top 1% DS1 results (in Table 6) to illustrate the processing that took place. Here, the AVG figures suggest that TANS has the best overall performance at the 1% cutoff, and with TAN, the conventional weighted form of the Tanimoto coefficient, performing least well. Table 11 shows that the value of the Kendall coefficient, 0.271, 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 six methods is TANS > BINS > BBNS > BBN > BIN > TAN.

The good performance of TANS is not restricted to this particular combination (top 1% for DS1), since it also the gives the best results for top 5% for DS1 and DS4. The poor performance of TAN is similarly not restricted, since it performs least well in six other combinations, that is, everything except both DS3 searches and top 5% for DS5. Indeed, the results obtained here are so consistently poor that one of the principal conclusions that can be drawn from the experiments is that use of raw fragment frequencies in the Tanimoto coefficient is to be avoided if effective screening is to be achieved. The results are better than if the simple binary form of the coefficient is employed[23] but are generally far inferior to those obtained when the square roots of the frequencies are employed.

The DS3 searches are of particular interest since they involve the most heterogeneous activity classes in the first four data sets and thus provide a stiff test of the effectiveness of a screening method. Chen et al. found that BIN was inferior to TAN for the DS3 activity classes. However, when the new network described here, BBN, is used on this data set, Tables 8 and 13 show that it (in the BBNS form) gives the best performance of all the methods for this data set at both cutoffs.

If ligand-based virtual screening 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 had been the inspiration for the design of the DS3 data set,[25] but the MUV data set has taken this idea much further. Specifically, each of the 17 sets of 30 PubChem actives in DS5 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 Bayesian network methods for searching structurally diverse sets of actives. The search results for DS5 are shown in Table 10, and are rather different from those for DS3. A network method is again the best, but this time it is BINS (top 1%) or BIN (top 5%), with BBNS and BBN yielding poor results, that is, a near-inversion of the ordering in Table 8.

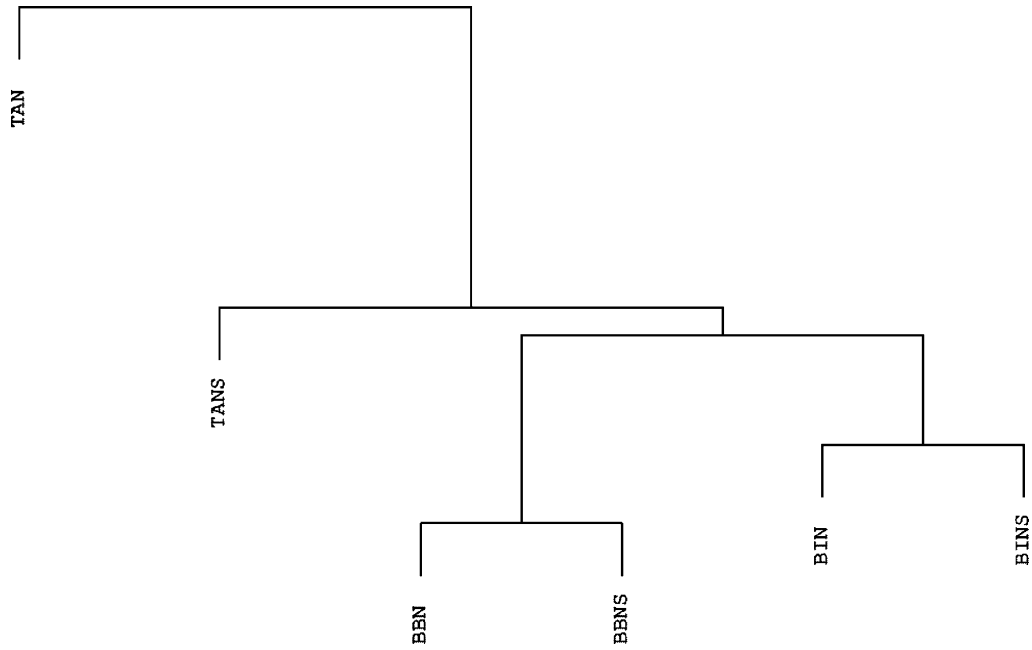

Figure 3. Complete linkage dendrogram showing the extent of the overlap between the top 1% outputs for searches of the DS2 database.

What, then, can we conclude from the results presented here? First, that there is no single best method: this is hardly surprising, since Sheridan, for example, has suggested that it will never be possible to identify some single similarity method that can offer a consistently high level of screening performance in all circumstances.[29] Second, taking the square root of the fragment frequencies can bring about substantial increases in performance in some cases. This has already been observed in the case of TAN as described above and is also the case (though for a lesser extent) for BIN; however, the increase in performance is negligible when this normalization is used with BBN. Thus, the mean ranks of the six methods over all of the ten rankings (five data sets and two cutoffs) are as follows: TANS = BINS (2.96) > BBNS (3.45) > BBN (3.46) > BIN (3.67) > TAN (4.45). A comparison of the computed similarity values for BBN and BBNS shows that there are often only slight differences between them, whereas the differences are much greater for BIN and BINS. This is because of the greater complexity of the BBN weighting function, which involves three factors that are not present in the BIN OKA weighting function and that are hence unaffected when the square root is used. Third, as we have noted previously and as these average ranks make clear, TANS is a highly effective screening method, especially when it is remembered that the processing is far simpler than for the two network models. Indeed, although the work reported here was undertaken to validate the utility of network models, our principal conclusion is that there is no obvious gain to be had from the use of these more complex approaches to screening, as compared to TANS: this is simple in both concept and implementation but still offers reasonably high levels of screening. Indeed, we would go further and suggest that TANS should be used as the baseline of performance when new types of similarity-based screening method are describes in the future; from the results obtained here and elsewhere,[23] we believe that the use of the conventional, binary Tanimoto coefficient provides too low a level of performance to provide a sufficiently stringent basis for comparative studies.

We further studied the relationship between the six similarity measures using an overlap analysis. Here, we looked at the numbers of molecules common to the outputs of pairs of searches. The results obtained are exemplified by Table 16, which lists the numbers of common molecules (averaged over all searches for all activity classes) in the top-1% outputs for the DS2 data set. The matrix is represented diagrammatically by the complete linkage dendrogram shown in Figure 3, which makes clear the very different nature of the TAN outputs from those resulting from the other five measures. Table 16 considers the top-ranked molecules irrespective of whether they were actives or inactives; a similar pattern of behavior is observed if attention is restricted to just the actives in the top 1% of the rankings. Very similar results to those for DS2 are obtained for the other data sets.

CONCLUSION

This paper has further investigated the use of Bayesian networks for ligand-based virtual screening. Experiments with MDDR, WOMBAT, and MUV data show that these approaches allow effective screening searches to be carried out. However, the results obtained are not obviously superior to those obtained using a much simpler approach that is based on the use of the Tanimoto coefficient and of the square roots of fragment occurrence frequencies.

REFERENCES AND NOTES

(1) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Kluwer: Dordrecht, The Netherlands, 2003.
(2) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein−ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **2006**, *49*, 5851–5855.
(3) Lengauer, T.; Lemmen, C.; Rarey, M.; Zimmermann, M. Novel technologies for virtual screening. *Drug Discovery Today* **2004**, *9*, 27–34.

(4) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903–911.
(5) Johnson, M. A.; Maggiora, G. M. *Concepts and Application of Molecular Similarity*; John Wiley: New York, 1990.
(6) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
(7) Nikolova, N.; Jaworska, J. Approaches to measure chemical similarity−A review. *QSAR Comb. Sci.* **2003**, *22*, 1006–1026.
(8) Bender, A.; Glen, R. C. Molecular similarity: A key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.
(9) Maldonado, A.; Doucet, J.; Petitjean, M.; Fan, B.-T. Molecular similarity and diversity in chemoinformatics: From theory to applications. *Mol. Diversity* **2006**, *10*, 39–79.
(10) 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, 1990; Association for Computing Machinery: New York, 1990; pp 1−24.
(11) Ribeiro, B.; Muntz, R. In *A Belief Network Model for IR*, Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, 1996; Association for Computing Machinery: New York, 1996; pp253−260.
(12) de Campos, L. M.; Fernández-Luna, J. M.; Huete, J. F. The BNR model: Foundations and performance of a Bayesian network-based retrieval model. *Int. J. Approximate Reasoning* **2003**, *34*, 265–285.
(13) Willett, P., Textual and chemical information retrieval: different applications but similar algorithms. *Inf. Research* [Online] **2000**, *5*, paper 69. http://informationr.net/ir/5-2/paper69.html (accessed April 20, 2010).
(14) Abdo, A.; Salim, N. In *Inference Networks for Chemical Similarity Searching*, Proceeding of the International Conference on Advanced Computer Theory and Engineering, Phuket, Thailand, 20−22 December, 2008; IEEE Computer Society: Los Alamitos, CA, 2008; pp 408−412.
(15) Abdo, A.; Salim, N. Similarity-based virtual screening with a Bayesian inference network. *ChemMedChem.* **2009**, *4*, 210–218.
(16) Abdo, A.; Salim, N. In *Molecular Similarity Searching Using Inference Network*, Proceeding of the 237th ACS National Meeting of the American Chemical Society, Salt Lake City, UT, 22−26 March, 2009; American Chemical Society: Washington, DC, 2009.
(17) Symyx Technologies. MDL Drug Data Report. http://www.symyx.com/products/databases/bioactivity/mddr/index.jsp (accessed April 20, 2010).
(18) Chen, B.; Mueller, C.; Willett, P., Evaluation of a Bayesian inference network for ligand-based virtual screening. *J. Cheminf.* **2009**, *1*, DOI: 10.1186/1758-2946-1-5. http://www.jcheminf.com/content/1/1/5 (accessed April 20, 2010).
(19) Sunset Molecular Discovery. World of Molecular Bioactivity. http://www.sunsetmolecular.com/ (accessed April 20, 2010).
(20) Turtle, H.; Croft, W. B. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.* **1991**, *9*, 187–222.
(21) Greiff, W. R.; Croft, W. B.; Turtle, H. PIC matrices: A computationally tractable class of probabilistic query operators. *ACM Trans. Inf. Syst.* **1999**, *17*, 367–405.
(22) 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.
(23) Arif, S.; Holliday, J.; Willett, P. Analysis and use of fragment-occurrence data in similarity-based virtual screening. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 655–668.
(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) Gardiner, E. J.; Gillet, V. J.; Haranczyk, M.; Hert, J.; Holliday, J. D.; Malim, N.; Patel, Y.; Willett, P. Turbo similarity searching: effect of fingerprint and dataset on virtual-screening performance. *Stat. Anal. Data Mining.* **2009**, *2*, 103–114.
(27) 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.
(28) Siegel, S.; Castellan, N. J. *Nonparametric Statistics for The Behavioral Sciences*; McGraw-Hill: New York, 1988.
(29) Sheridan, R. P. Chemical similarity searches: When is complexity justified. *Expert Opin. Drug Discovery* **2007**, *2*, 423–430.

CI100090P